信号の位相特性に着目した
音源情報知覚に関する研究

Perception of Sound Source Information
based on
the Phase Spectrum

2010年 2月

早稲田大学大学院国際情報通信研究科
国際情報通信学専攻 音響情報処理研究II

後藤　理

概要

周波数分析により求められる信号情報の多くは振幅特性により表現される。本論文では軽視されがちな位相特性が狭帯域包絡線を通して音源情報として知覚されることを明らかにし、位相特性と音源情報知覚の関係について論ずる。

聴覚モデルにおいて音響信号は蝸牛の周波数選択性から狭帯域信号に分割される。また、分割された狭帯域信号は有毛細胞モデルにより半波整流後にローパスフィルターを施した狭帯域包絡線を得ると考えられている。Meddis 等 (1997) はこのようにして得られた狭帯域包絡線の自己相関関数を用いた Summary Autocorrelation Function (SACF) による聴覚脳幹の周期性検出に基づく Pitch 知覚モデルを提案した。さらに、狭帯域包絡線は音声の了解性などの音源情報知覚にも寄与していることが知られている。Drullman 等 (1997) は 100Hz-6.4kHz にわたる 24 帯域の 1/4 オクターブ帯域包絡線とそれぞれの周波数帯域に対応する帯域雑音から了解性のある音声を合成できることを報告し、Shannon 等 (1995) はさらに概ね 4 帯域にわたる帯域包絡線を用いて了解性のある音声が実現できることを示した。このように、狭帯域包絡線を保存した合成信号により了解性のある音声が得られることが知られている。このように、狭帯域包絡線により音源情報を保存し知覚可能である。

また、音声の了解性において Schroeder 等 (1986), と Traumueller 等 (1987) は位相情報より母音が復元可能であることを示唆した。また、Oppenheim 等 (1981) は信号長が十分長いとき振幅のみの音声復元では音声の了解性は失われ、位相情報では失われなかったことを報告した。これらの研究は音源情報の

特徴表現にあまり使用されない位相特性に音声の了解性に関わる情報が含まれていることを示唆している。

本論文ではこのような狭帯域包絡線と音声了解性の関係に着目し、位相特性に着目した音声合成実験を行い、2ms 以下のごく短いフレーム長と 256ms 以上の長いフレーム長の短時間フーリエ変換により合成した音声信号では、位相特性を保存することにより了解性のある音声が復元できることを明らかにする。さらに、それらのフレーム長において了解性とともに情報知覚に重要な狭帯域包絡線が復元されることを示す。これは、周波数分析窓長における振幅・位相情報の優位性を示すとともに、音源情報の知覚と狭帯域包絡線の関係を改めて示すものである。このような知見を元に、音源情報の類似性、言語情報と同様に重要な話者情報と知覚の関係を狭帯域包絡線の入れ替えを通じた合成信号の試聴実験を元に論じ、その結果、音源情報知覚において位相情報に音源情報が含まれ、その位相情報が狭帯域包絡線を通し知覚されることを明らかにする。

目 次

図 目 次

第1章　序論

1.1　本研究の意義と目的

本研究は音声における音声了解性、音節明瞭度のような言語情報と話者情報を音源情報とし、音源情報と知覚の関係について位相情報を通じ論じる。音源情報は音源の特定、音源分離、構造物診断、音場の把握、ヒアリングエイドなど数多くの分野で利用される。このような音源情報は分析において物理特徴と関連した振幅周波数特性分析が多く用いられている。本論文は音源情報を、振幅周波数特性と対を成す位相特性の変化に見いだし、位相特性が音源情報知覚さらには音源識別に及ぼす影響効果について論じるものである。本論文では音源特徴を含む位相特性の変化が狭帯域包絡線を通して知覚されると考える。本論文において位相特性の変化と知覚の関係は、音声了解性試聴実験の分析において位相情報による狭帯域包絡線復元を示すことにより結ばれる。さらに、音声の代表的な客観評価値である音声了解度、音声明瞭度、マスキング効果における信号類似度、話者識別に関するそれぞれの試聴実験を通し、狭帯域包絡線に音源情報が含まれることが示される。本研究の成果は、いままで明らかにされていない位相特性と音源情報知覚の関係を、情報知覚に重要であると考えられる狭帯域包絡線に着目し明らかにすることにある。また、狭帯域包絡線が帯域分割による周波数情報と包絡線による時間変化情報を含むことから、聴覚における時間－周波数分析の関係に言及する。さらに狭帯域包絡線に含まれる音源情報の所在として音声了解性に関わる情報は狭帯域包絡線類似度、音質に関

わる情報は狭帯域包絡線の振幅ヒストグラム、話者特徴は狭帯域包絡線の帯域間類似度に現れることを示す。また、これらの音源情報に着目し実用的な客観評価法をそれぞれ提案する。

1.2　本研究分野に関する背景

音に含まれる情報は、信号における情報知覚の容易さと関連し、音声了解度や信号品質の評価、さらには音源情報そのものを取り扱う、コミュニケーションエイドや音の可視化等に利用されている。

音源情報の分析におて、短時間フーリエ変換を用いたスペクトル分析が挙げられる。短時間フーリエ変換により信号は振幅スペクトルと位相スペクトルに分析される。それらのスペクトル特徴から特に振幅スペクトルが音声信号の分析・合成に用いられてきた [Schroeder, 1999]。また、振幅スペクトルの構成に現れる音素間の差とパワースペクトルサブトラクションが音源情報を保存する雑音除去技術に使われてきた [Boll, 1979; Vary, 1985]。一方、位相特性に関する研究は位相情報を用いて母音が復元可能であることを Schroeder 等 [Schroeder and Strube, 1986] , と Traumueller 等 [Traumueller and Schouten, 1987] は示唆している。また、Oppenheim 等 [Oppenheim and Lim, 1981] は信号長が十分長いとき振幅のみの音声復元では音声の了解性は失われ、位相情報では失われなかったことを報告した。これらの研究は位相特性に音声の了解性に関わる情報が含まれていることを示唆している。

音源情報と関わる音声了解性において狭帯域包絡線が音源情報知覚に寄与していることが知られている。Drullman [Drullman, 1995] は 100Hz - 6.4kHz にわたる 24 帯域の 1/4 オクターブ帯域包絡線とそれぞれの周波数帯域に対応する帯域雑音から了解性のある音声を合成できることを報告した。Shannon 等 [Shannon et al., 1995] はさらに概ね 4 帯域にわたる帯域包絡線を用いて了解

性のある音声が実現できることを示した。このように、狭帯域包絡線を保存した合成信号により了解性のある音声が得られることが報告されている。本論文では了解性において重要とされる狭帯域包絡線を位相特性による復元を試み、その関係を明らかにする。さらに本論文では位相特性の変化が狭帯域包絡線を通して知覚されると考え、音源情報知覚とその関係を音声における代表的な客観評価値により考察する。音声了解性と共に音声信号において代表的な情報として信号の品質情報が挙げられる。本論文では、位相情報により構成される狭帯域包絡線に着目し信号の品質評価を試みる。また、人間が類似する信号が重畳された音を知覚するとき音源情報の理解と分離が難しくなる情報マスキングが知られている [Freyman et al., 1999] , [Arbogast et al., 2002] , [Schmitz and Iyer, 2003] 。この情報マスキング評価は信号の類似性による評価が行われている。本論文では、情報マスキング効果による試聴実験を通じ狭帯域包絡線と信号の類似度の知覚について考察する。さらに、音声信号において了解性と同様に重要な話者情報において、Li 等 [Li and Hughes, 1974] はスペクトルの時間変化による話者特徴分析により狭帯域包絡線に含まれる話者情報に言及した。しかし、狭帯域包絡線と話者情報の知覚の関係は示されていない。本論文では合成信号の試聴実験を通じ、狭帯域包絡線と話者情報知覚について言及する。本論文は以上のような代表的な音声に関する情報を位相特性に着目し狭帯域包絡線を通じて客観評価を行い、その情報の所在を明らかにするものである。さらにその音源情報をもとに実用的な評価手法を示すことを試みる。

1.3 本研究分野に関する従来技術

本論文は試聴実験を通して物理特徴と知覚の関係を考察する。知覚と関わる音源情報を数値化し評価する試みは続けられている。音声了解性に関わる音源情報評価において、Houtgast 等 [Houtgast and Steeneken, 1973] は音声の包

絡線変化に着目して音声了解度を予測する MTF-STI 法 (Modulation Transfer Function - Speech Transmission Index) を提案している。さらに、音声了解性と共に音声信号において代表的な情報として信号の品質情報が挙げられる。音声信号の品質評価には、通信における劣化信号に対し PESQ(Perceptual Evaluation of Speech Quality) [Rix et al., 2001] [Beerends et al., 2002] を用いた評価が行われている。この PESQ 値は通信におけるパケットロスや周波数歪みを考慮し総合的な音声品質を音声品質評価における Opinion 値の予測として算出する。さらに、音響信号の質として信号の類似度がある。人間が類似する信号が重畳された音を知覚するとき音源情報の理解と分離が難しくなる情報マスキングが知られている [Freyman et al., 1999] , [Arbogast et al., 2002] , [Schmitz and Iyer, 2003]。この情報マスキング評価は信号の類似性による評価が行われている。さらに、話者特徴において MFCC(Mel-Frequency Cepstrum Coffiecient) による GMM(Gaussian Mixture Model) [Campbell, 1997] を用いた話者識別が行われている。本研究は、音源情報と知覚を論じるとともに、これらの従来技術に対し、新しい評価手法を提案する。

1.4　本論文の構成

第 1 章は、序論として本研究の背景及び目的・意義について述べる.

第 2 章では、音声了解度における短時間フーリエ変換のフレーム長と位相情報の関係について述べる。試聴実験において 1/16ms - 2048ms のフレーム長を用いた音声信号の振幅情報と位相情報を白色雑音による振幅位相情報と入れ替えた合成信号により音声了解度を調べ、256ms より長いフレーム長と 4ms より短いフレーム長において位相情報が音声了解度を保存することを明らかにする。

第 3 章では、了解性のある音声信号と位相情報が関連することから、狭帯域

包絡線と位相情報の関係について考察する。本章では位相スペクトルの自己相関関数により狭帯域包絡線周波数を推定する。さらに、音声明瞭度試聴実験を行い、位相スペクトルより求めた狭帯域包絡線周波数から音声明瞭度評価を試みる。その結果、信号の位相特性により音声了解性が評価可能であることを示唆し、位相特性と狭帯域包絡線の関係を示す。さらに従来法として S/N と関わる MTF - STI と位相情報による音声明瞭度評価の比較を行い、位相情報を用いた音情報表現の可能性を示す。

第 4 章では、音源情報の総合的な情報に関わる信号の品質評価について狭帯域包絡線相関係数を用いた予測法を考察する。本章では音声品質評価において従来法とされる PESQ 値と狭帯域包絡線相関係数による信号音質評価予測の比較を行う。その結果、実環境による評価試験では Opinion 評価値が低い場合 PESQ による予測が困難となることに対し、狭帯域包絡線の振幅ヒストグラム歪みでは予測可能なことを示し、狭帯域包絡線振幅ヒストグラムの歪みによる信号音質評価予測の可能性を示す。

第 5 章では、音源情報と狭帯域包絡線の関係を信号の類似性に基づき考察する。狭帯域包絡線に音源情報が含まれる場合、信号情報に基づく類似性は時間波形における振幅変化で評価できると考えた。そこで試聴実験において情報マスキング効果に着目した信号類似度評価をおこない、信号の振幅分布から信号の類似度が評価可能であることを示し、狭帯域包絡線に音声信号以外の音源特徴も含まれることを明らかにする。

第 6 章では、狭帯域包絡線に音源特徴が含まれることから、狭帯域包絡線と音声に含まれる話者情報の関係について考察する。Li 等 Li and Hughes [1974] はスペクトルの時間変化による話者特徴分析により狭帯域包絡線に含まれる話者情報に言及した。しかし、狭帯域包絡線と話者情報の知覚の関係は示されていない。そこで音声信号における狭帯域包絡線と狭帯域搬送波を入れ替えた合成

信号による話者判定試聴実験を行い狭帯域包絡線に話者情報が含まれていることを明らかにする。

第 7 章では、第 6 章で明らかにした狭帯域包絡線に含まれる話者情報から、帯域間相関行列による話者特徴表現を試みる。さらに、従来法における MFCC(Mel-Frequency Cepstrum Coffiecient) による GMM(Gaussian Mixture Model) を用いた話者識別結果と比較する。その結果、従来法である MFCC による GMM を用いた話者識別に対して帯域間相関行列を用いた話者識別では識別率が向上することを示し、帯域間相関行列による話者特徴表現の可能性を示す。

第 8 章では、本論文の総括を述べる。

第2章　位相スペクトルと音声了解性

2.1　まえがき

本章は短時間フーリエスペクトルの位相情報に着目し音声の了解性と位相スペクトルの関係について述べる。音響信号は短時間フーリエ変換により振幅スペクトルと位相スペクトルに分析される。それらのスペクトル特徴から特に振幅スペクトルが音響信号の分析・合成に用いられてきた [Schroeder, 1999]。ホルマントに代表される音声分析・合成はその顕著な例である。一方、室内音場における信号処理では音の位相スペクトル変化が長く継続する残響音の効果を表現する上で重要であろうと考えられてきた。本章では信号分析・合成に用いられる振幅スペクトルに対し、位相スペクトルと音情報知覚の関係を明らかにする。

2.2　位相スペクトルと音声知覚

Schroeder 等 [Schroeder and Strube, 1986] , と Traumueller 等 [Traumueller and Schouten, 1987] は長い時間にわたって観測した位相情報より母音が復元可能であること、また Oppenheim 等 [Oppenheim and Lim, 1981] は信号長が十分長いとき振幅スペクトルのみに着目したの音声復元では音声の了解性が失われ、了解性を回復するには位相情報が重要であることを報告している。しかし了解性のある音声合成における位相情報の重要性に関する試聴実験評価はこ

れまで報告がなく、現在も明確ではない。

Liu 等 [Liu et al., 1997] は VCV 音声信号の母音間の閉鎖子音の知覚に対する位相の効果を研究している。そのことにより、母音間の閉鎖子音の知覚がフーリエ変換の分析フレーム長により変化し、フレーム長が長くなるほど振幅から位相情報が重要になっていき 192ms と 256ms の間で振幅位相スペクトルの優位性が交代することを報告している。またごく短い窓長において位相変化が知覚されることも観測されている。

また、調波位置の検出が主に振幅情報に基づいて行われるのに対し、閉鎖子音の認知は位相情報を手掛かりにして行われるものと結論付けている。

一方、音声の了解性は狭帯域包絡線と関係していると考えられている。Drullman [Drullman, 1995] は了解性のある音声信号が 24 個の 1/4oct. 帯域 (100-6400Hz) の狭帯域包絡線によって変調した帯域別変調雑音から合成できることを発見した。本論文では人は位相情報に表れる変化を信号の帯域別包絡線を通じて知覚すると考える。本章では発話文章の了解性を位相情報と帯域別包絡線の両面から分析し、位相情報にともなう音声の了解性の変化を考察する。その結果、信号分析時間長による位相スペクトルの優位性の変化を帯域別包絡線の復元に着目して明らかにする。この信号分析長の変化に存在する了解性の変化は人間が音声を知覚する際の時間－周波数分析機構の特徴を示すものであると考えられる。

2.3　合成音声信号による試聴実験

合成（振幅もしくは位相のみ）音声信号は女声の発話と雑音を用いそのサンプルを図 2.1 に示す。フレーム長を変えた短時間フーリエ変換分析による 2 つの合成信号の文章了解度は試聴実験によって評価した。

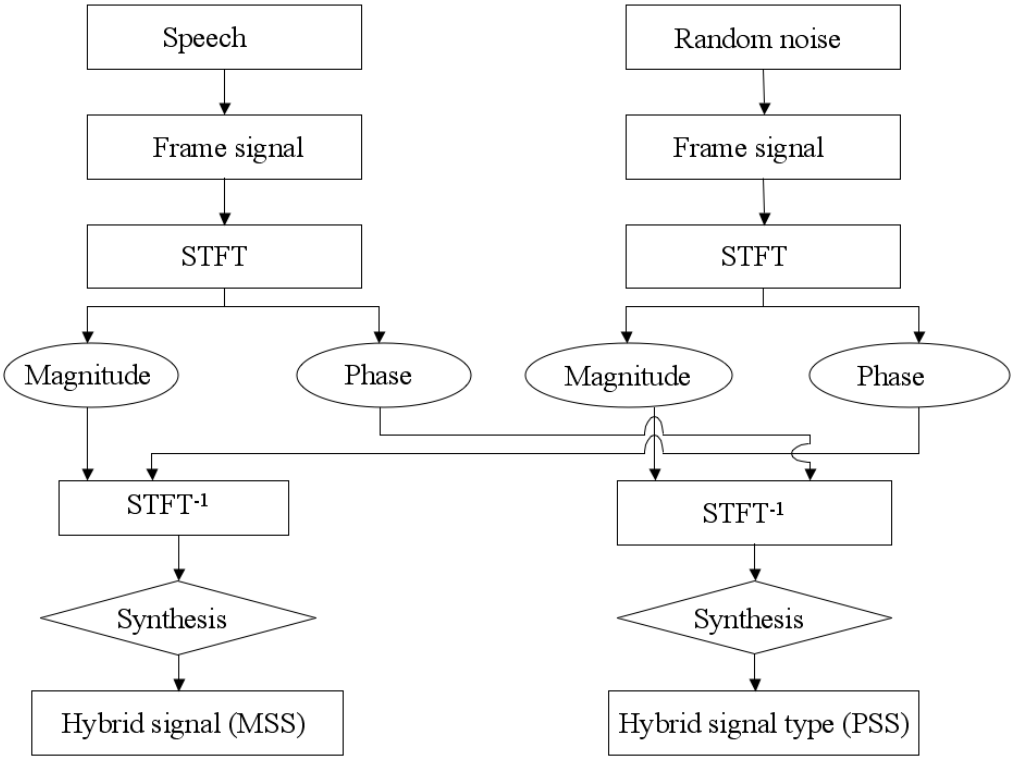

図 2.1: 短時間フーリエ変換を用いた合成信号作成方法

2.3.1　実験方法

被験者と音声信号

被験者は年齢２４－４９の７名でいずれも日本語を母国語とする健聴者である。原音声信号は２名の女声による日本語音声９６発話を用いた。信号の長さは１．５秒の音声信号の先頭と末尾に無音区間を加えた４秒である。音声文章は日常的な文章で６～１０の単語を組み合わせたものとした（例：この字は遠くから見えにくい）。また白色雑音は MATLAB(プログラム関数名:randn) により生成した。

信号処理

音声と雑音の組は互いにオーバーラップする三角窓により信号を切り出し短時間フーリエ変換により分析した。ただし、フレーム長が１ポイントと２ポイントにおいてオーバーラップは無いものとした。図 2.1 に示されている通り２つの合成信号が合成される。合成信号 MSS は音声の振幅スペクトルと雑音の位相スペクトルから合成され、反対に PSS は音声の位相スペクトルと雑音の振幅スペクトルから合成される。合成においても分析と同様に三角窓を利用した。分析合成に用いたフレームの長さは 1/16ms から 2048ms にわたる 16 種類とした。全てのフレーム長に各 6 文章を使用した。

試験方法

被験者は 192 個 (96 個の MSS と 96 個の PSS) の合成信号を試聴する。合成信号は１条件につき 6 個である。ランダムに再生される 192 個の合成信号を被験者はヘッドホンを通して（ダイオティック）好みのレベルで再生し聴いた通りに文章を書き取った。書き取った文章が原音声文章に完全に一致した場合の

み正解とし、全試聴文章中における正解文章の割合を了解度とした。了解度はMSS もしくは PSS ともにフレーム長ごとに集計した。

2.3.2 結果

図 2.2 に文章了解度の結果を示す。各データは被験者にあたり 6 個の試験結果の平均値である。文章了解度 100 %は全ての被験者が正しく文章を書き取れたことを示す。横軸はフレームの長さを示す。MSS 合成信号の了解度はフレーム長による明らかな変化を示した。4-64ms にわたる中程度のフレーム長ではほぼ完全に近い文章了解度を示し、反対に短いフレーム長と長いフレーム長では完全に文章了解度を失っている。PSS 合成信号は MSS と対照的な傾向を示しているところが極めて興味深い。

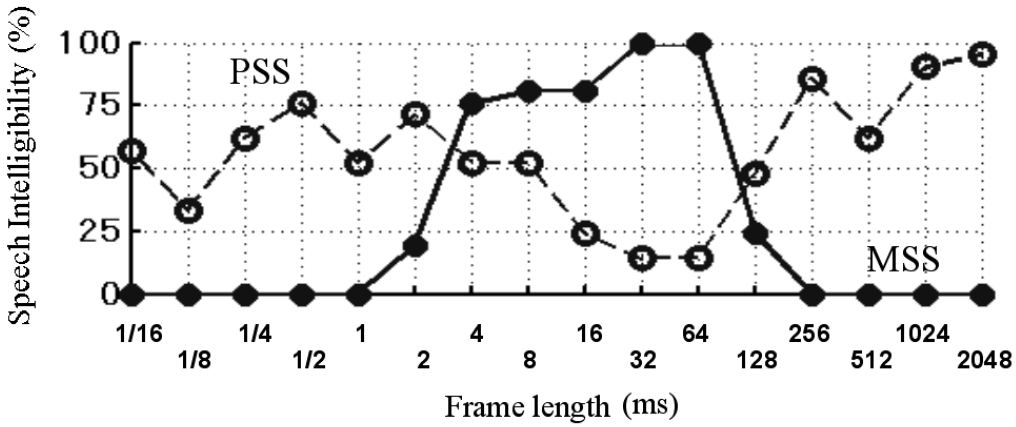

図 2.2: 短時間フーリエ変換におけるフレーム長と位相スペクトルによる合成音声 (PSS) と振幅スペクトルによる合成音声 (MSS) の音声了解度

2.4　狭帯域包絡線の保存と音声了解度

まえがきに述べたとおり Drullman [Drullman, 1995] は狭帯域包絡線が音声の了解性につながる事を明らかにしている。そこで本節においては 2 種類の合成信号 MSS と PSS における狭帯域包絡線の変化を考察する事とした。

2.4.1　合成信号の狭帯域包絡線の観測

図 2.3 に原信号と合成信号 MSS と PSS の帯域別包絡線の例を描いた。図は 1 帯域（1kHz 1/4 オクターブ帯域）における 3 種類のフレーム長（1/2、32、2048ms）の例である。

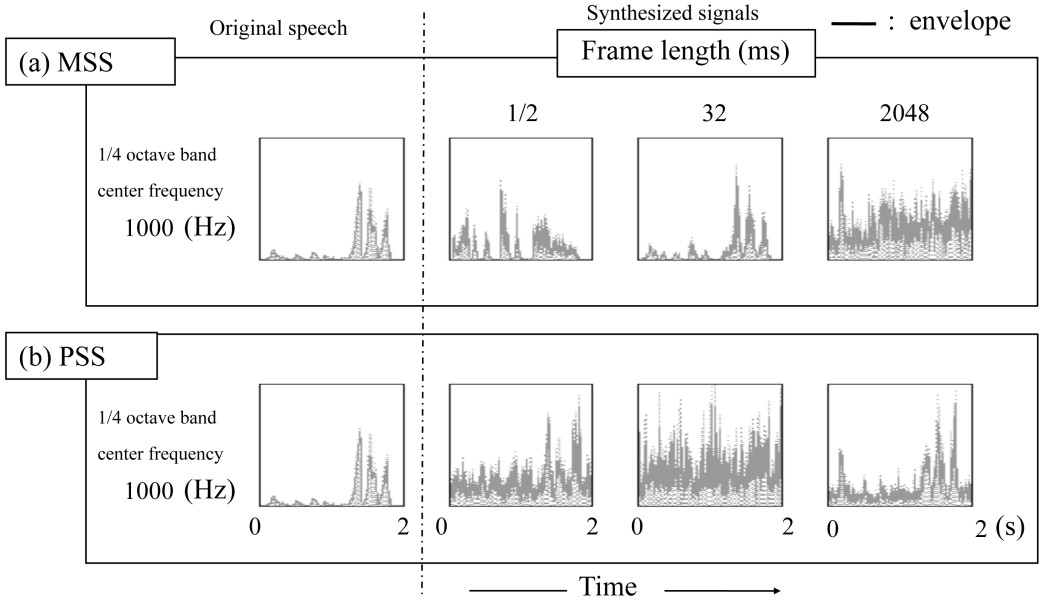

図 2.3: 原信号と合成信号 MSS と PSS の帯域別包絡線 (1/4 オクターブ帯域 fc:1kHz)

図 2.3a は MSS による狭帯域包絡線の例である。図より 32ms のフレーム長

のみ原信号の包絡線と似た形となることがわかる。しかし図 2.3b に示した PSS による包絡線例は図 3a の MSS と逆の傾向 (32ms 以外において原信号の包絡線と似た形が復元されている) を示していることがわかる。

このように包絡線の観測と了解度試験結果の関係に着目した事から本章ではフレーム長の変化による狭帯域包絡線の復元度を狭帯域包絡線相関係数によって考察する事にした。

2.4.2 狭帯域包絡線相関係数による包絡線復元分析

狭帯域包絡線の相関係数は 1/4 オクターブ帯域毎に計算した。番号 i 番目の帯域における文章 l の相関係数を

$$\rho_i(l) = \frac{\overline{\hat{e}_{oi}(l,n)\hat{e}_{si}(l,n)}}{\sqrt{\hat{E}_{oi}(l)\hat{E}_{si}(l)}} \tag{2.1}$$

と定義する。ここで

$$\hat{e}_{oi}(l,n) = e_{oi}(l,n) - \overline{e_{oi}(l,n)} \tag{2.2}$$

$$\hat{e}_{si}(l,n) = e_{si}(l,n) - \overline{e_{si}(l,n)} \tag{2.3}$$

$$\hat{E}_{oi}(l) = \overline{\hat{e}_{oi}(l,n)^2} \tag{2.4}$$

$$\hat{E}_{si}(l) = \overline{\hat{e}_{si}(l,n)^2} \tag{2.5}$$

とする。但し $e_{oi}(l,n)$ と $e_{si}(l,n)$ は i 番目の帯域の l 番目の文章の合成信号の自乗包絡線を、$\overline{(*)}$ は時間平均を示す。相関係数はさらに了解度試験に使用された全 96 文による集合平均によって求めた。

図 2.4b-e は合成信号と原信号の 1/4oct. 帯域毎の相関係数の例を示す。図 2.4a は図 2.2 の了解度試験の結果を再度示したものである。時間窓長を変化させた

相関係数は周波数帯域に依存することが見てとれる。さらに MSS と PSS の相関係数の間に見られる相補的性質が観測される。図に見るとおり了解度試験結果と狭帯域包絡線相関は分析フレーム長との関係においてほぼ同様の傾向を示していることが解る。この了解性と狭帯域包絡線相関係数の対応から、狭帯域包絡線の保存と音声了解性が深く関わっていることを確認できる。

上図において MSS と PSS による相関係数は 2 カ所で交差している。縦の破線で示した約 256ms のフレーム長における交差点は周波数帯域によらず一致する。この MSS と PSS に対する了解性あるいは包絡線相関係数がフレーム長 256ms で交差する背景には音声の支配的な包絡線変調周波数が関係していると思われる。相関係数の変化に関わる遅い変化を伴う音声包絡線は、了解度の変化に寄与している事が考えられるであろう。

一方、縦の点線で見られる交差は周波数に依存する。即ち交差するフレーム長はそれぞれの帯域中心周波数の周期と概ね一致する。これは振幅スペクトルから狭帯域包絡線を復元するには周波数帯域に関係する最低のフレーム長が必要となることを示唆している。以上の実験観察結果をまとめると以下のとおりとなる。

(1)MSS に対する結果は常識的である。即ち、長いフレーム長 (>256ms) においては包絡線を追従復元する時間分解能が十分である。反対により短いフレーム長 (<4ms) に関して、周波数分解能が不十分になる (帯域の中心周波数によると考えられる)。

(2) 一方 PSS は興味深い結果を示している。即ち包絡線は 256ms より長いフレーム長とごく短いフレーム長において (部分的でも) 復元する。

以下、本論文では位相が支配的になる長いフレーム長とごく短いフレーム長における位相スペクトルによる狭帯域包絡線の復元についてさらに考察をする。

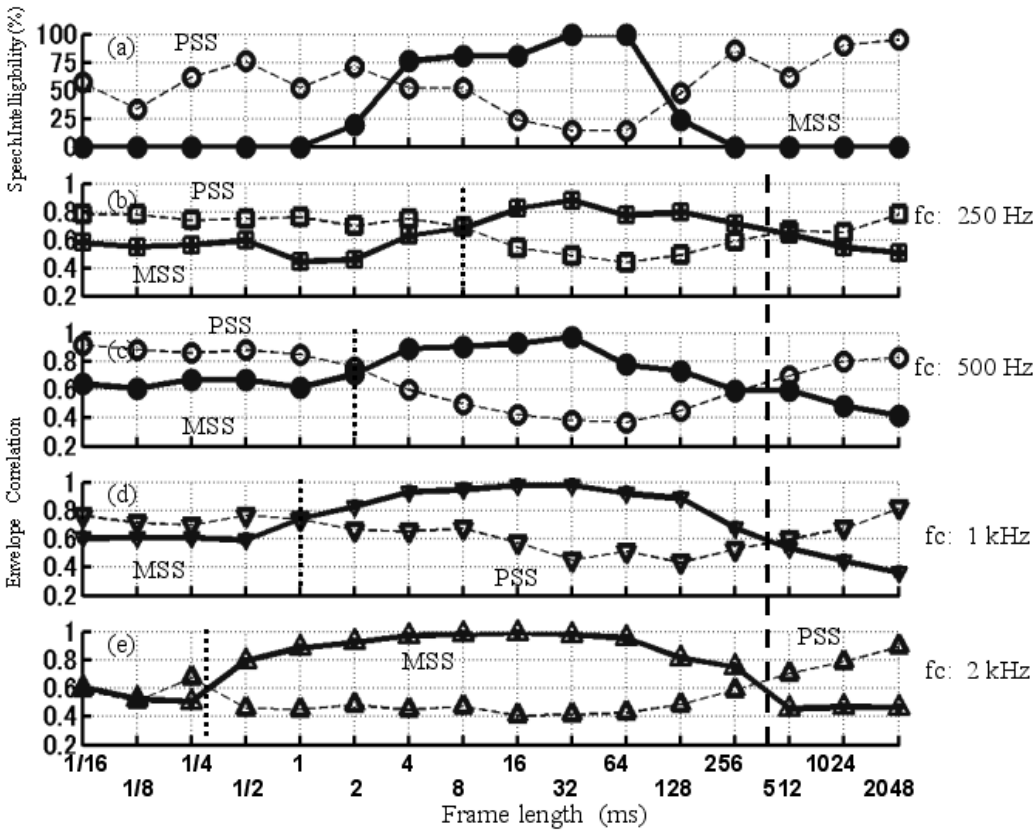

図 2.4: 音声了解度 (a) と 合成信号と原信号の狭帯域包絡線相関係数 (b-e)

2.4.3　位相スペクトルによる狭帯域包絡線の復元

長いフレーム長における位相スペクトルの重要性

図 2.5 は (Aa) に定常雑音 (Ba) に余弦波変調雑音 (b) に振幅スペクトル (c) に位相スペクトルをそれぞれ描く。この例における包絡線変調周波数は

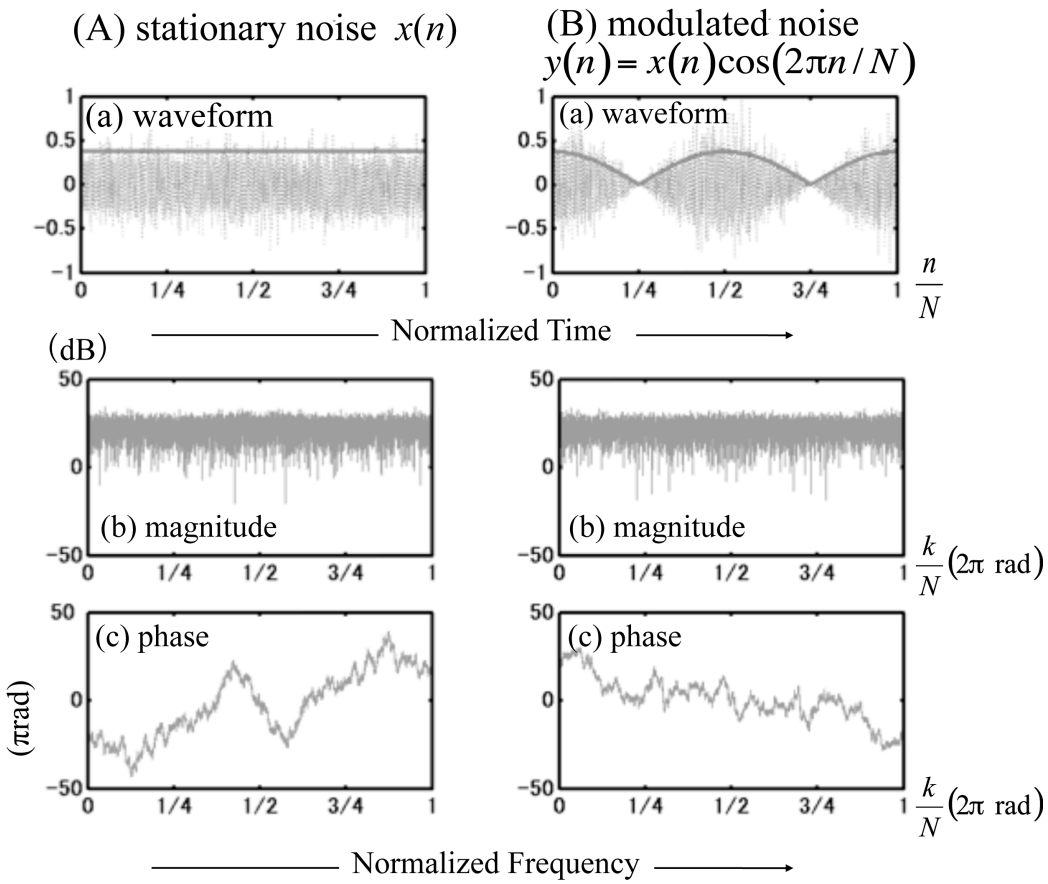

図 2.5: (A) 定常雑音 (B) 変調雑音 (a) 信号波形 (b) 振幅スペクトル (c) 位相スペクトル

$$2k_0 = 2(1/N) \tag{2.6}$$

で得られ N は信号長を表す。ここにおける短時間フーリエ分析は全信号長を用い計算した。しかし振幅スペクトルならびに位相スペクトルいずれからも包

絡線周波数の手掛かりを示すことができない。そこで位相スペクトルを用いた相関分析による包絡線周波数の観測を行った。図 2.6 にその結果を示す。

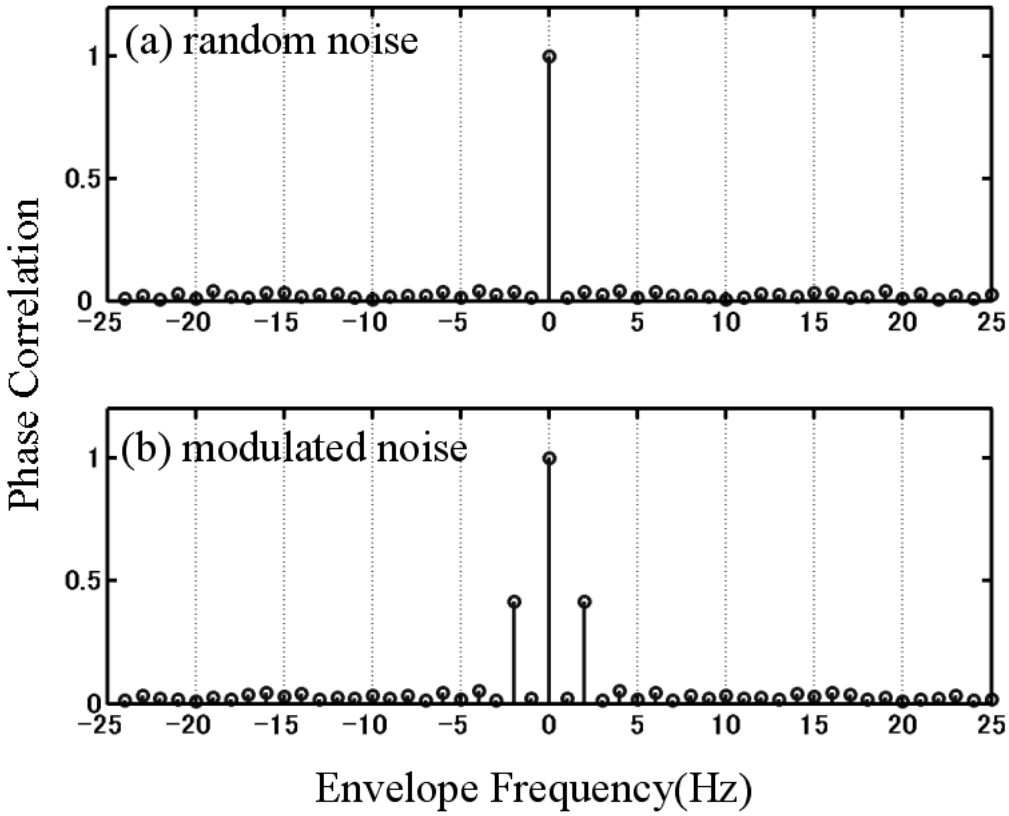

図 2.6: 図 2.5 における信号を用いた位相相関分析

周波数 (k') と $(k+k')$ における信号の位相差を

$$\Delta\theta(k,k') = \theta(k+k') - \theta(k') \tag{2.7}$$

として、位相相関関数 $phc(k)$ を

$$phc_c(k) = \frac{1}{K}\sum_{k'=0}^{k'=K-1} \cos\Delta\theta(k,k') \tag{2.8}$$

$$phc_s(k) = \frac{1}{K}\sum_{k'=0}^{k'=K-1} \sin\Delta\theta(k,k') \tag{2.9}$$

$$phc(k) = \sqrt{phc_c(k)^2 + phc_s(k)^2}. \tag{2.10}$$

と定義する。ここで K は対象となる周波数要素の数を表す。図において横軸は包絡線周波数と対応する。図 2.6b に示した変調雑音の位相相関関数分析より変調周波数が位相スペクトルから推定できることが読みとれる。図 2.7 は図

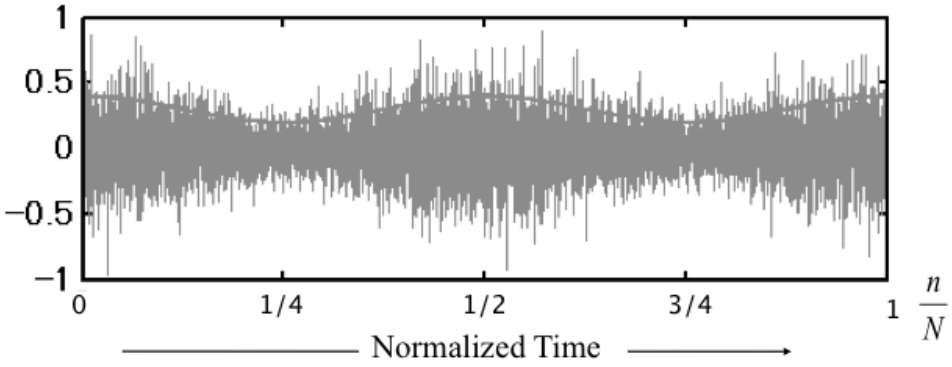

図 2.7: 2.5B における変調雑音の位相スペクトルによる合成

2.5b に示した変調雑音の振幅スペクトルを白色雑音の振幅と入れ替えた合成信号の例を示す。これにより位相スペクトルに包絡線に関する情報が一部保存されていることが分かる。しかし位相スペクトルから包絡線を復元するには位相スペクトルを分析する周波数分解能、即ち包絡線の周期より長い短時間フーリエ変換のフレーム長が必要であることも明らかになった。図 2.4 で見た位相からの包絡線復元のフレーム長は 256ms より長い必要がある。この結果は音声の主要包絡線周波数は 4Hz 周辺にあることを示唆している。

ごく短いフレーム長における位相情報の重要性

2.3.2 節に述べたごく短いフレーム長に関わる位相情報の優位性は、波形の零交差情報による狭帯域包絡線の復元として解釈できる。既に、図 2.4b-e に見たとおり包絡線の復元にはフレーム長の中心周波数に関連する周期より短いフレーム長であることが重要なことを示していた。

そこで短いフレーム長の限界として 1/16ms のサンプル１点による短時間フーリエ変換を考えてみる。1 点短時間フーリエ変換による各サンプルにおける振幅と位相は、各サンプルの瞬時振幅と正か負かの符号となる。このように 1 点短時間フーリエ変換による位相情報はサンプリング周波数が適正であれば零交差情報を保存している。零交差情報は振幅情報こそ失われているが波形の符号変化を保存している信号である。このような零交差情報を保存する信号は原信号の狭帯域包絡線を部分的に復元できる。

図 2.8 と図 2.9 は振幅スペクトル情報が部分的に零交差情報から復元される例を示すものである。図 2.8a は定常雑音波形、図 2.8b は符号変化を保存し振幅を 1 に置き換えて合成した 1 点短時間フーリエ変換による雑音の例である。図 2.8c と図 2.8d は上記２つの雑音の帯域別エネルギーの時間変化である。これらの図より原信号の低い周波数帯域における振幅スペクトルは位相情報のみを用いた 1 点短時間フーリエ変換による信号合成により保存されることがわかる。図 2.8d に見られる高域の周波数帯域におけるエネルギーの増大は、図 2.8b に示されるような波形の過度な振幅クリッピングによる雑音だと解釈できる。

図 2.8 と同様に図 2.9 は変調された帯域を持つ雑音の一点フーリエ分析による例である。図 2.9d に示すように原信号の振幅情報を失っているにもかかわらず一点フーリエ分析の位相情報 (零交差の情報) から狭帯域包絡線が部分的に復元することがわかる。

上記は図 2.4 で確認したごく短いフレーム長における位相の重要性を説明す

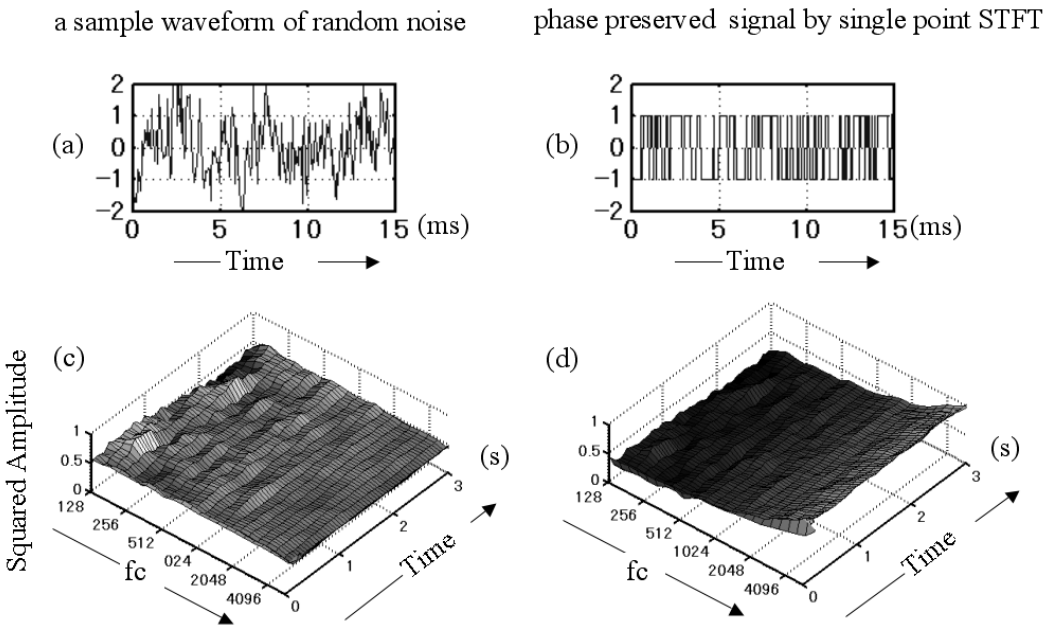

図 2.8: 定常雑音の 1 点短時間フーリエ変換による位相保存信号例

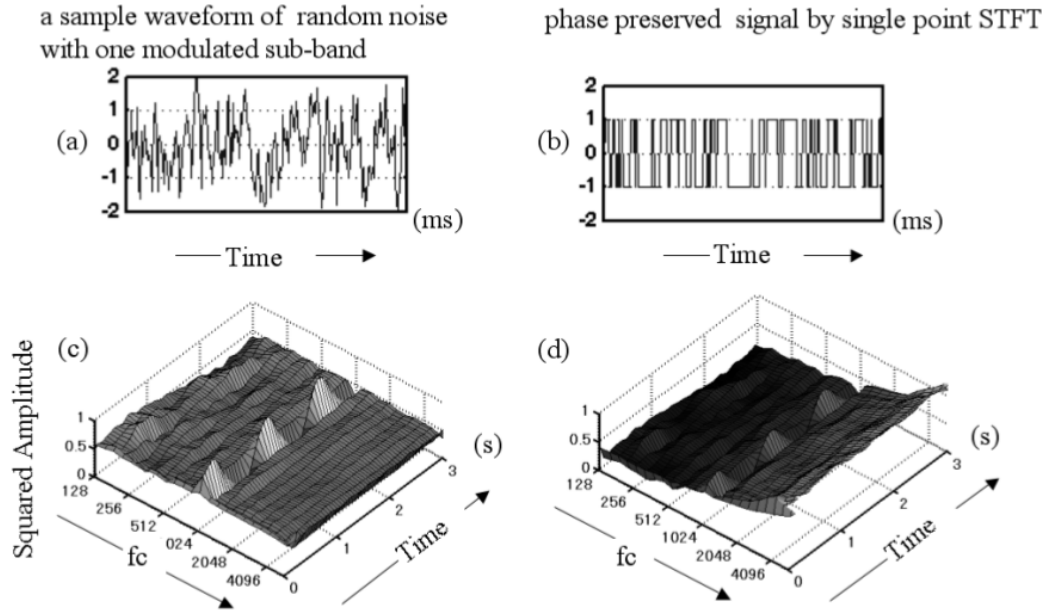

図 2.9: 変調帯域をもつ雑音の 1 点短時間フーリエ変換による位相保存信号例

るものである。しかし一方で図 2.4 は高域の包絡線はごく短いフレーム長における位相情報では復元出来ないことも示している。これは音声全体のパワースペクトルで見ればごく一部だけが高域に分布する。したがって高域の変調周波数特徴は信号の零交差の統計にはわずかにしか含まれない事を示唆していると思われる。

2.5 考察

本章では音声了解性に関わる位相情報の重要性について考察した。実験において音声の振幅 (位相) スペクトルと雑音による位相 (振幅) による合成音声信号の了解性を調べた。その結果、4-64ms のフレーム長において音声の振幅スペクトルと雑音の位相スペクトルを用いて了解性のある合成信号が得られることがわかった。一方 256ms より長いフレーム長と 4ms より短いフレーム長において了解性のある合成音声には音声の位相スペクトルと雑音の振幅スペクトルが重要である事がわかった。

音声了解性試聴実験の結果を合成信号と原信号の 1/4oct. 帯域における包絡線の相互相関係数を比較した結果、フレーム長の変化にともない了解度と相関係数が同じ傾向を示した。この傾向の一致から狭帯域包絡線が音声了解性の保存に関して重要な要素である事を確認できた。

音声の振幅スペクトルを用いた音声合成において、フレーム長が音声包絡線の支配的な周期 (約 256ms) より長くなると時間分解能を失い合成音声は了解性を失うこととなる。また、フレーム長が 4ms より短い場合、合成音声は対応する周波数分解能が不十分となり了解性を失う。試聴実験の結果は音声の振幅スペクトルを用いた合成信号が了解性を失うフレーム長において、反対に位相スペクトルを用いた合成信号が了解性を持つことを示した。

長いフレーム長における位相スペクトルは位相スペクトルの自己相関分析で

示したように包絡線情報を含んでいる。また、ごく短いフレーム長における位相スペクトルのみの合成信号は原信号のパワースペクトルから得られるような零交差間隔情報を持っている。これは各帯域の狭帯域包絡線が全体域信号の零交差波から一部復元されることからも確認できる。このように振幅スペクトルのフレーム長によって時間 (もしくは周波数) 解像度が失われると位相スペクトルの周波数 (もしくは時間) 解像度が代替する。これは、長いフレーム長において位相の周波数分解能が狭帯域包絡線を復元し、またごく短いフレーム長において位相スペクトルの時間分解能が狭帯域包絡線を一部復元できる零交差情報を保存できる事を示している。

この研究の結果、音声了解性と狭帯域包絡線の保存における短時間フーリエ変換のフレーム長変化について、振幅情報が優位になる中程度のフレーム長に対して２つの位相が優位になる領域があることがわかった。さらに、それぞれのフレーム長において位相情報から狭帯域包絡線が復元できることから、位相情報の変化は狭帯域包絡線を通じて知覚していることが考えられる。

2.6　むすび

本章では、音声了解性と位相スペクトルの関係について短時間フーリエ変換のフレーム長変化における合成信号試聴実験により明らかにした。その結果、音声了解性の保存において振幅情報が優位になる中程度のフレーム長に対し、ごく短いフレーム長と長いフレーム長の２つの位相スペクトルが優位になる領域があることがわかった。さらに、それぞれのフレーム長において位相情報から狭帯域包絡線が復元できることを示した。これより位相スペクトルと音情報知覚に関する音声了解性の関係を示した。

第3章　位相相関関数を用いた音声の包絡線表現と明瞭度

3.1　まえがき

前章に述べたとおり音声信号に十分な長さがあればフーリエ変換による振幅スペクトルから合成した音声は了解性を失うのに対して、位相スペクトルを用いて合成した音声には了解性がある。さらに短いフレーム長においても位相情報が音の知覚に重要であることが明らかとなった。一方、Drullman [Drullman, 1995] は狭帯域音声の波形包絡線が了解性のある音声を表現するには重要であることを述べている。また Houtgast 等 [Houtgast and Steeneken, 1973] は音声の包絡線変化に着目して音声了解度を予測する MTF-STI 法 (Modulation Transfer Function - Speech Transmission Index) を提案している。

音声の了解度あるいは明瞭度は、人が音声を含む複数の音源信号の中から音声を知覚する機能を表す最も基本的な評価値である。音声信号に重畳する音源の単純な例では、音声信号に無相関な定常雑音が代表的な例である。従来より雑音を含む音声の聴き取り評価では、雑音と音声のエネルギー (S/N) が着目されてきた。そこで本章では S/N に代わって前章に述べた位相相関 (PCI) に着目して、定常雑音が重畳する音声の単音節明瞭度の予測評価を試みる。その結果従来の S/N に伴って変化する包絡線変化に着目した MTF-STI による予測結果と比較検討する。位相変化に基づいて MTF-STI 法と同様に明瞭度を予測することができるとすれば、それは人は位相情報を包絡線の変化として知覚して

いることを示唆するものであろう。

3.2　雑音下における音声明瞭度試験

本章では前章に述べた文章了解度に代わり、より細かな音声了解性と関わる音節明瞭度に着目する。従来の音節明瞭度試験は試験音節の前後にキャリアフレーズを配置して、個別に発声された音節ごとに聴き取り試験を行うものであった。本実験では音場における明瞭度試験により適していると思われる連続発声された無意味 3 音節を続けて聞き取る試験を行うこととした。雑音中の音声明瞭度試験音の収音には無響室を利用した。図 3.1 は無響室内スピーカ配置である。互いに無相関の広帯域定常雑音を６つのスピーカから再生する。音声

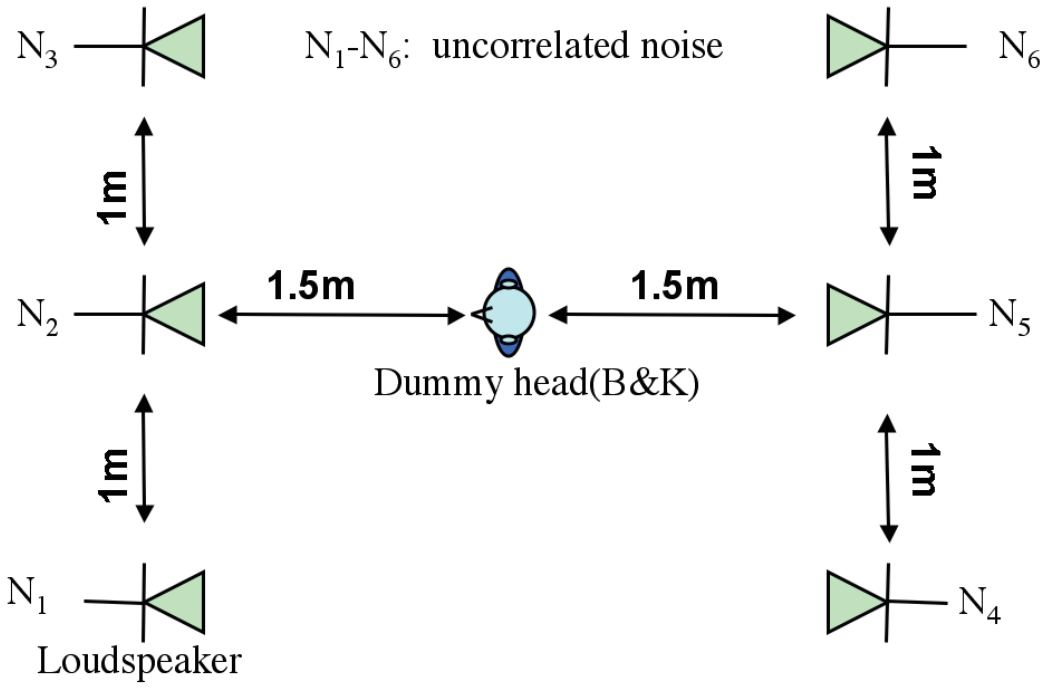

図 3.1: 試験音収音図

信号となる無意味３音節はこの６つのスピーカの中から無作為に選んだ１つのスピーカより再生する。無意味３音節音声全体と雑音の平均 S/N は 0 もしく

は30dBの2種類とした。試験音(3音節)の総数は150サンプルとし、その中で120をS/N 0 dB、残り30を30dBとした。試験音あたりの信号の長さは、無意味3音節の前後にそれぞれ無声区間1.5秒を付加した全長約4秒である。図3.1中の中央においてダミーヘッド(Bruel and Koer)を用いて(サンプリング周波数44.1kHz、量子化ビット数16bitのA/Dコンバータ)収音した。

被験者は5人の健聴者男性である。被験者は150サンプルの収録された試験音をダイコティック受聴で好みの音量にて受聴し書き取った。音節明瞭度は第２音節の正答率で評価した。これは本実験がキャリアフレーズを含まない再生方式であることから、被験者が第１音節を聴き逃す場合が無視できないことによる。

3.3 位相相関分析

前章で述べた位相相関分析に基づいて位相相関数列PCSを改めて定義する。

$$PCS(k) \equiv \sqrt{S^2(k) + C^2(k)} \tag{3.1}$$

$$S(k) \equiv \frac{1}{N}\sum_{l=0}^{N-1} \sin \Delta\theta(k,l) \tag{3.2}$$

$$C(k) \equiv \frac{1}{N}\sum_{l=0}^{N-1} \cos \Delta\theta(k,l) \tag{3.3}$$

$$\Delta\theta(k,l) \equiv \theta(l+k) - \theta(l) \tag{3.4}$$

ここで$\theta(k)$は信号$x(n)$の位相スペクトル、kは周波数、lは周波数シフト数を表す。

図3.2は試験音として収音した雑音下における狭帯域波形の例を示す。図3.2AaとAcは収音した無音声区間の雑音、図3.2Abは雑音を含む無意味3音

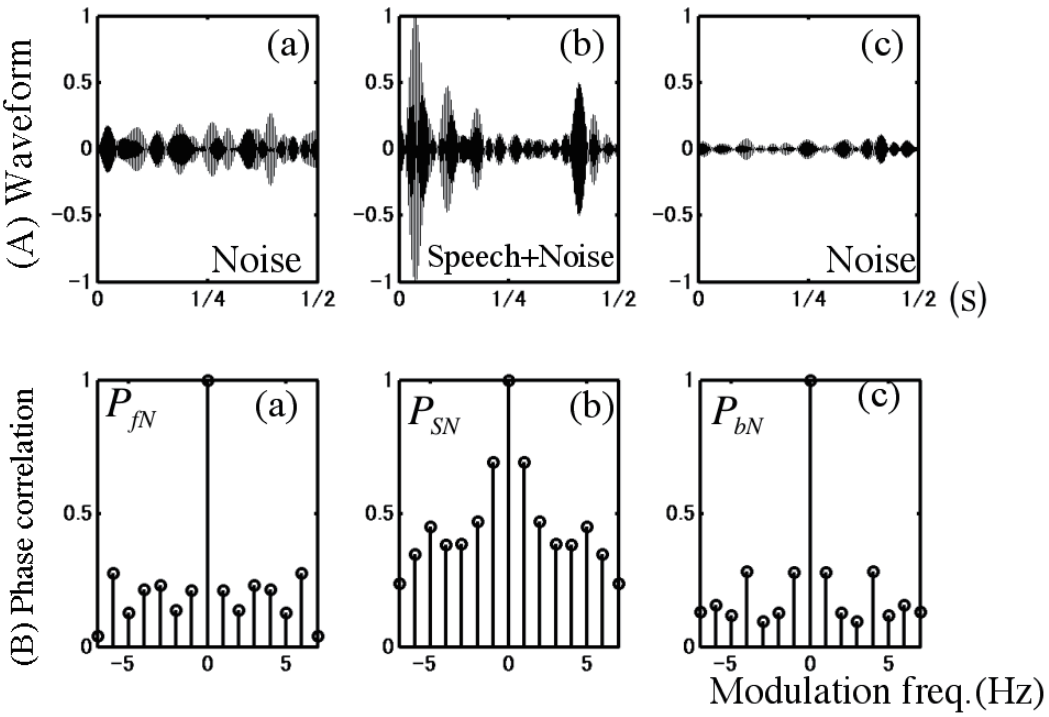

図 3.2: 試験音の狭帯域波形と PCS (1/4 オクターブ帯域 fc: 250(Hz))

節の例である。図 3.2Ba-c は図 3.2Aa-c より算出した $PCS_{fn}(k)$、$PCS_{SN}(k)$、$PCS_{bn}(k)$ をそれぞれ示す。

図 3.2 において試験音の明瞭度が高ければ音声区間が際立つことから、$PCS_{fn}(k)$ と $PCS_{SN}(k)$、あるいは $PCS_{SN}(k)$ と $PCS_{bn}(k)$ の差が大きくなることが見込まれる。そこで音節明瞭度を予測指数として PCI(Phase Correlation Index) を試験音ごとに

$$PCI(i) \equiv \frac{1}{M}\sum_{m=1}^{M} w(m)PCI(i,m) \tag{3.5}$$

$$PCI(i,m) \equiv 10\log\frac{D(i,m)}{D(m)} \tag{3.6}$$

$$D(i,m) \equiv \sum_{k=0}^{N-1} B^2(i,m) + F^2(i,m) \tag{3.7}$$

$$B(i,m) \equiv PCS_{SN}(i,m) - PCS_{bN}(i,m) \tag{3.8}$$

$$F(i,m) \equiv PCS_{SN}(i,m) - PCS_{fN}(i,m) \tag{3.9}$$

$$D(m) \equiv \frac{1}{Q}\sum_{i=1}^{Q} D(i,m) \tag{3.10}$$

と定義する。ここで i は試験音の番号、Q はサンプル数 (150 個) と M は周波数帯域の数である。本実験の周波数帯域は 250 ー 4700Hz の 1/4oct. 帯域による 17 帯域とした。また周波数帯域に渡る平均値を算出するに必要な重み関数を $w(m)$ とする。但し本実験では単純化のため $w(m) = 1$ とした。

図 3.3 は音声明瞭度と PCI による音声明瞭度予測結果である。横軸は PCI の値である。上式のとおり PCI は全試験音声に渡る平均値ではなく試験音ごとに求めるものである。即ち下部の棒グラフ (NTS(Number of Test Samples)) は横軸が示す PCI の値を有する試験音の数を示す。また同一試験音に対して被験者はダイコティック受聴によって左右異なる信号を受聴することから、PCI の値も二つの値が得られることになる。しかしここでは従来のベストイヤー仮説に従って左右を比べて品質が高いと思われる値 (PCI の大きさが大きい値) を選んだ。図よりサンプル数が少なくばらつきが大きくなる明瞭度が極めて高い場合を除いて PCI と音声明瞭度の間に高い相関を見ることができる。

本節では PCI における音声明瞭度評価値の信頼性を従来法である MTF-STI(Modulation Transfer Function - Speech Transmission Index) を用いて評価することとした。MTF-STI は図 3.4 [Houtgast et al., 1980] に示すように

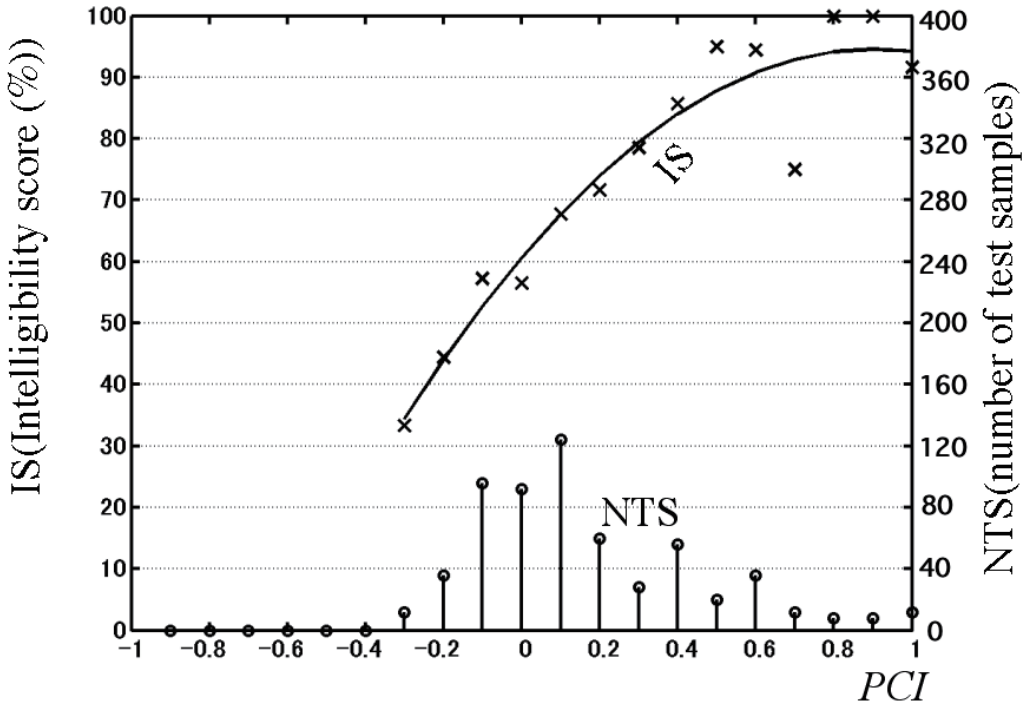

図 3.3: PCI と音声明瞭度

音声信号に雑音が加わった時の帯域包絡線の変調度を算出し計算する。さらに残響などの伝達関数による狭帯域包絡線の変調度の変化も計算することにより雑音と伝達系の影響を受けた音声明瞭度を予測することができる。

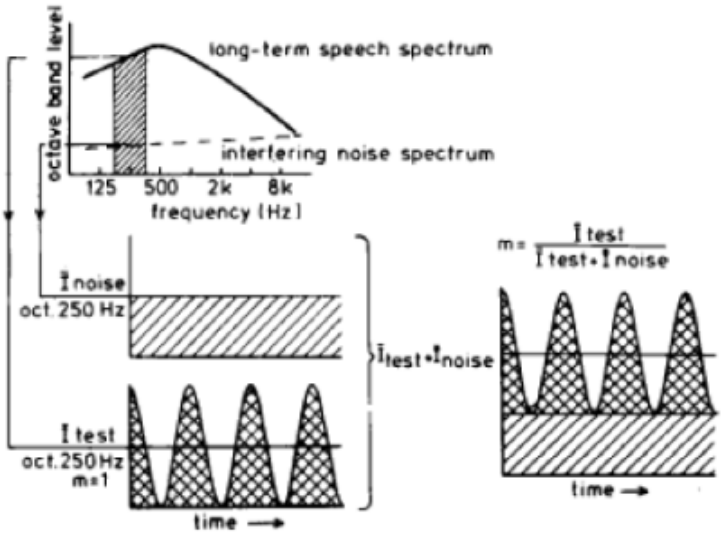

図 3.4: Modulation Transfer Function

図 3.5 には MTF-STI による音声明瞭度予測結果を示す。

この結果、広く用いられる MTF-STI が明瞭度評価に適正であるとすれば試聴実験結果の信頼性が高い事が確認できる。さらに、PCI と MTF-STI の相関が高いことも確認できた。これにより前章において位相情報と音声了解度の関係を示したことに対し、音声了解性が低い場合の評価に用いる音声明瞭度も位相情報から評価可能であることがわかった。

3.4 考察

本章では前章において明らかとなった音声了解性と位相情報の関係を、位相情報のみを用いた単音節明瞭度予測実験により考察した。その結果、位相情報のみを用いて音声明瞭度が予測可能なことを明らかにした。さらに従来法とし

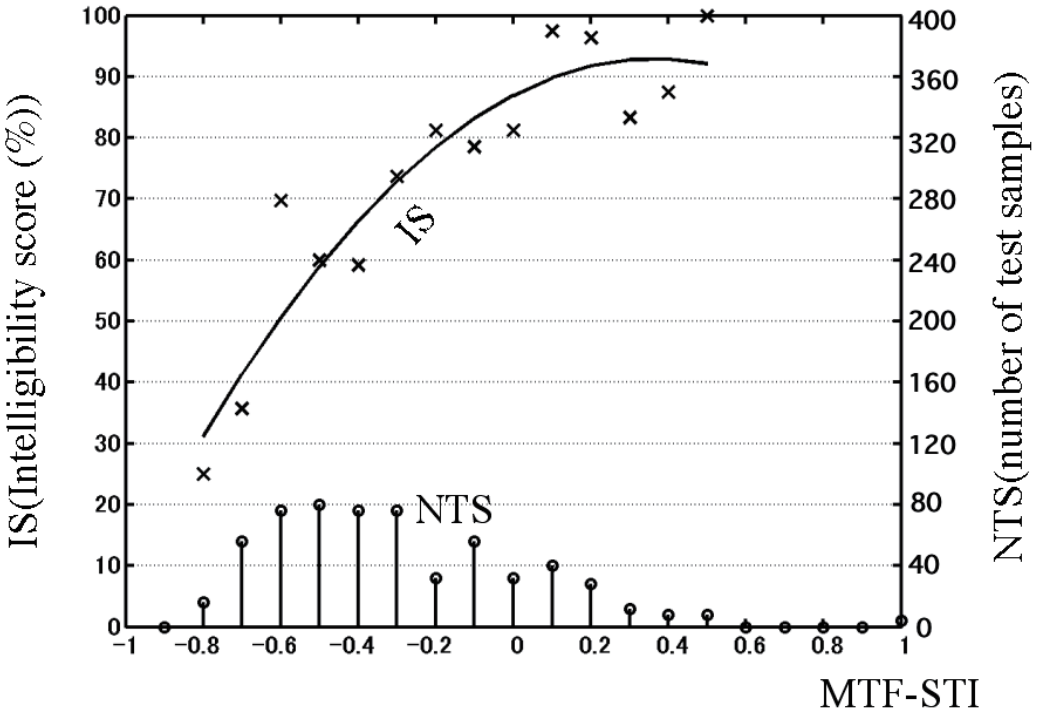

図 3.5: MTF-STI と音声明瞭度

て S/N と関わる MTF-STI と PCI による予測方法の比較を行い、互いの音声明瞭度予測が高い相関関係にあることを示した。これにより、従来の S/N と関わる振幅スペクトルによる音源情報の表現に対し、位相情報を用いた音情報表現の可能性を示した。

3.5 むすび

本章では S/N に代わって前章に述べた位相相関 (PCI) に着目して、定常雑音が重畳する音声の単音節明瞭度の予測評価を試みた。その結果、S/N に伴って変化する包絡線変化に着目した MTF-STI 法による予測結果と、位相相関 (PCI) を用いた単音節明瞭度予測結果の相関が高く、位相情報により MTF-STI 法と同様に明瞭度を予測できることがわかった。この結果より位相変化に基づいて明瞭度を予測可能なことから、人は位相情報を包絡線の変化として知覚していることが考えられる。これより、信号の振幅スペクトルと関係する S/N からではなく、位相情報により音声の知覚と関係する音声明瞭度が評価可能であることを試聴実験を元に明らかにした。

第4章　包絡線振幅ヒストグラムと音声品質オピニオン評価値

4.1　まえがき

前章において単音節明瞭度が位相情報を用いて予測評価可能なことを明らかにし、位相変化と知覚の関係に言及した。本章では、音声了解度が低い場合に評価可能な単音節明瞭度に対し、音声了解度が高い場合に評価に用いるopinion 評価による音声信号品質評価の予測を行う。

従来より音声の明瞭度、文章了解度 の予測については多くの研究が行われてきた。客観評価と呼ばれる明瞭度あるいは了解度の予測には、電話通話音声を対象として研究されてきた音節明瞭度予測指数 AI (Articulation Index) 法、また電話通話に限らず伝達系の音声了解度を評価する MTF-STI (Modulation Transfer Function - Speech Transmission Index) 法がよく知られている。一方、最近では通信網の発達により単純な音声の了解性評価ではない、通信による時間ひずみや通信機器によるエコーキャンセラーの影響を考慮した音声信号の品質評価が行われるようになってきた。本章では、このような音声品質評価をOpinion 評価実験を行い狭帯域包絡線に着目し予測を試みる。

4.2　音声信号の品質評価

音声明瞭度が高い場合、音声の内容はほぼ完全に聞き取れることから音声信号の評価に品質評価を用いる。本節ではOpinion 評価値を用い音声信号の品質

評価実験を行い音声明瞭度が高い場合における評価を行った。図 4.1 に Opinion 評価に用いた試験信号の収音状況の写真を示す。図に示すように中央にマイクロホン (図中矢印) を配置し 6 つのスピーカを用い前章における音声明瞭度評価と同様に試験音を収音した。また、Opinion 評価試験は指向性雑音と全指向

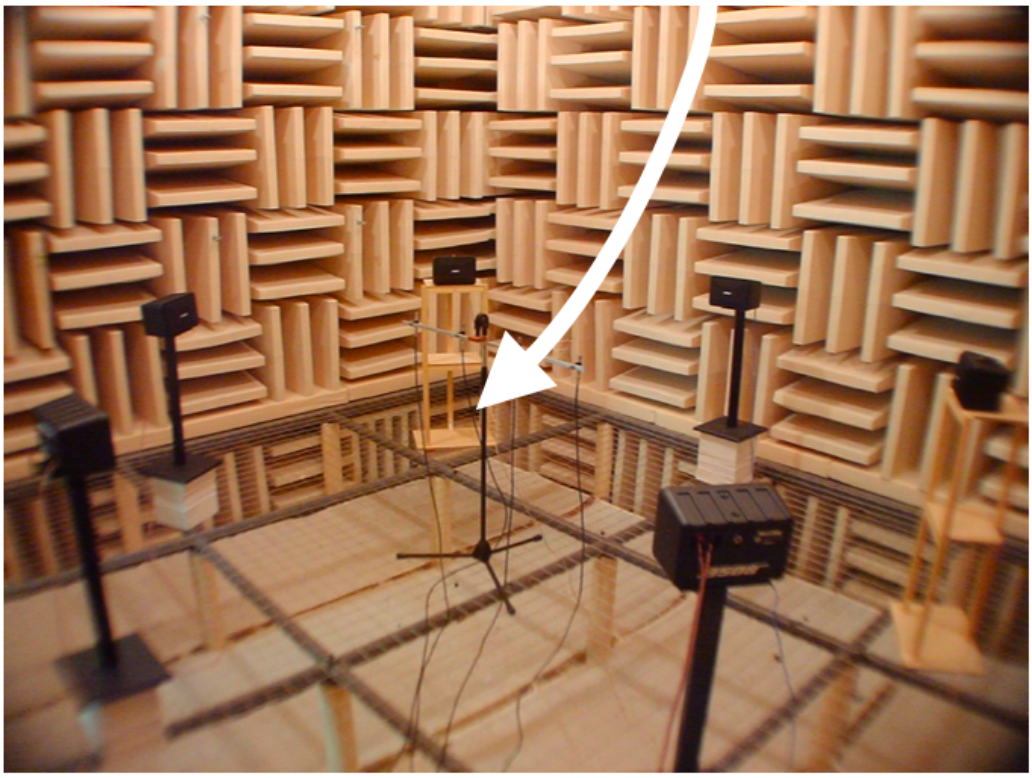

図 4.1: 無響室における収音風景

性雑音を用いた試験信号を用意し、図 4.2 と図 4.3 にそれぞれの収音状況を示す。マイクロホンは１点、２点、４点 (十字配置) の三条件とし、2 点と４点のマイクロホンによる収音信号は単純な Delay Sum により指向性収音とした。また、再生した試験信号には約 3 秒の音素バランス文章 (ATR) を使用しサンプリング周波数 44.1kHz、量子化ビット数 16bit の A/D コンバータを用いて収音した。それぞれ試験音の S/N は-3,0,3,6,9(dB) とし、試験音は各条件に 5 サン

プル (合計 75 サンプル) 作成した。被験者は 4 名の健聴者で提示信号の品質を「良い」～「悪い」の 5 段階で評価させ平均したものを Opinion 評価値とした。被験者は 5 人の健聴者の男性を用いて行った。

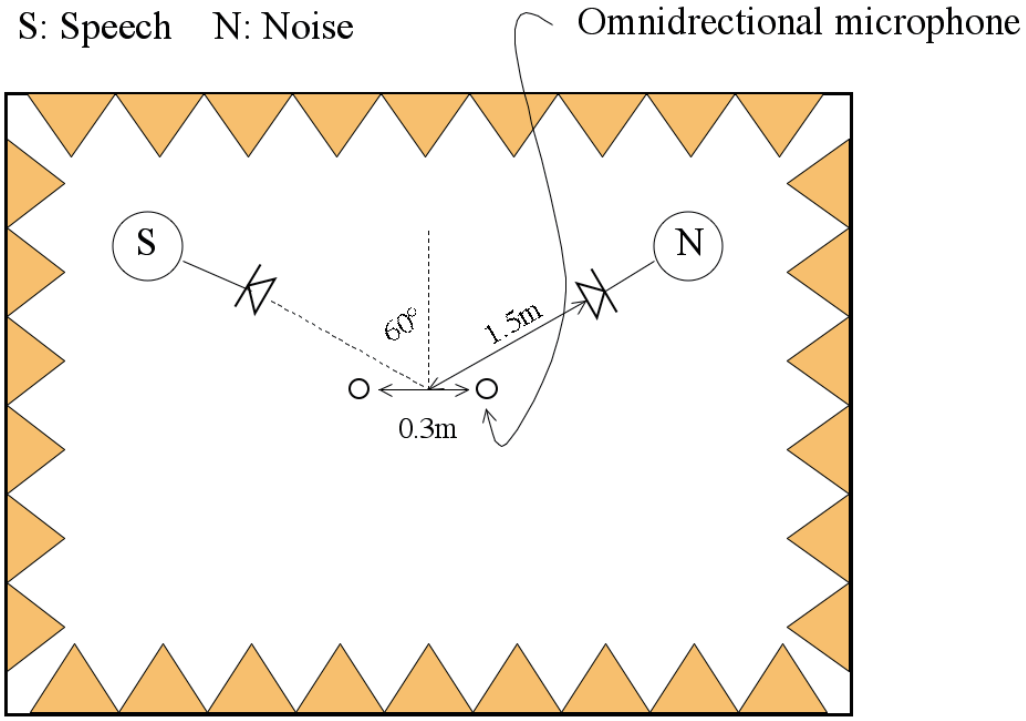

図 4.2: 指向性雑音による試験音収音状況

4.3 Opinion 評価値の予測

Opinion 評価値の予測に PESQ(Perceptual Evaluation of Speech Quality) [Rix et al., 2001] [Beerends et al., 2002] が知られている。図 4.4 に PESQ の算出法を示す。図に示すとおり PESQ 値は通信におけるパケットロスなどを考慮した波形の時間ずれや帯域別スペクトル歪みやマスキング効果等を考慮する Opinion 評価方法である。このような PESQ を用いた Opinion 評価値の予測結果を図 4.5 図より S/N の向上とマイクロホンの増加に対し PESQ 値も上昇し Opinion

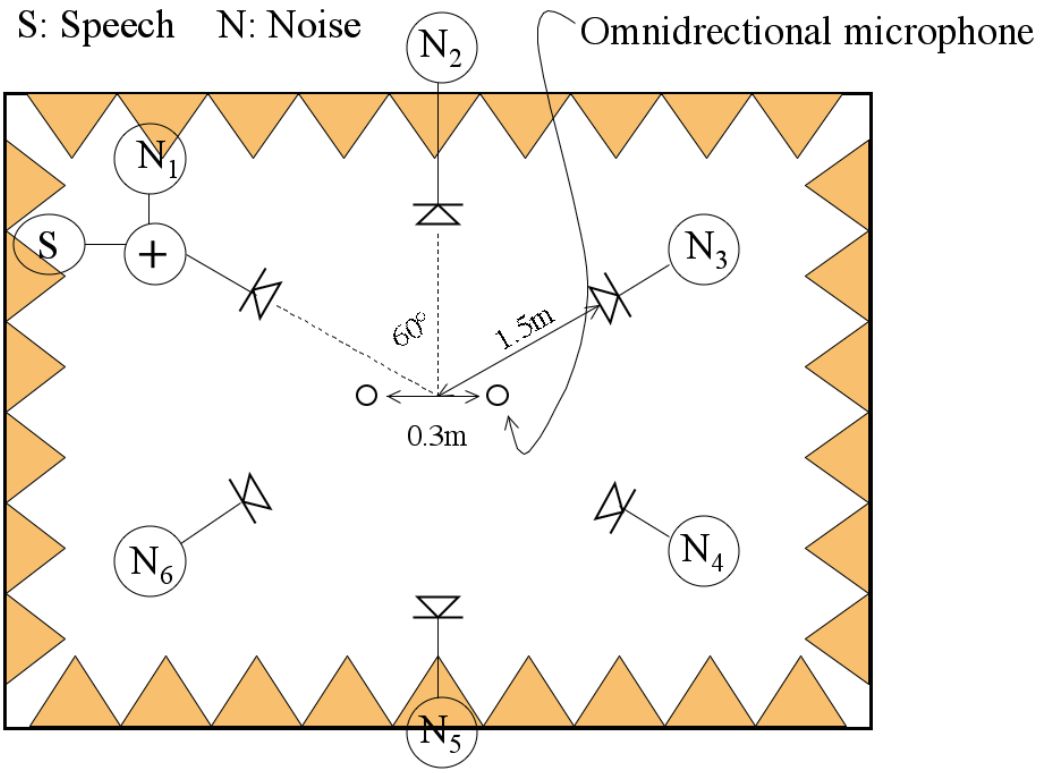

図 4.3: 全指向性雑音による試験音収音状況

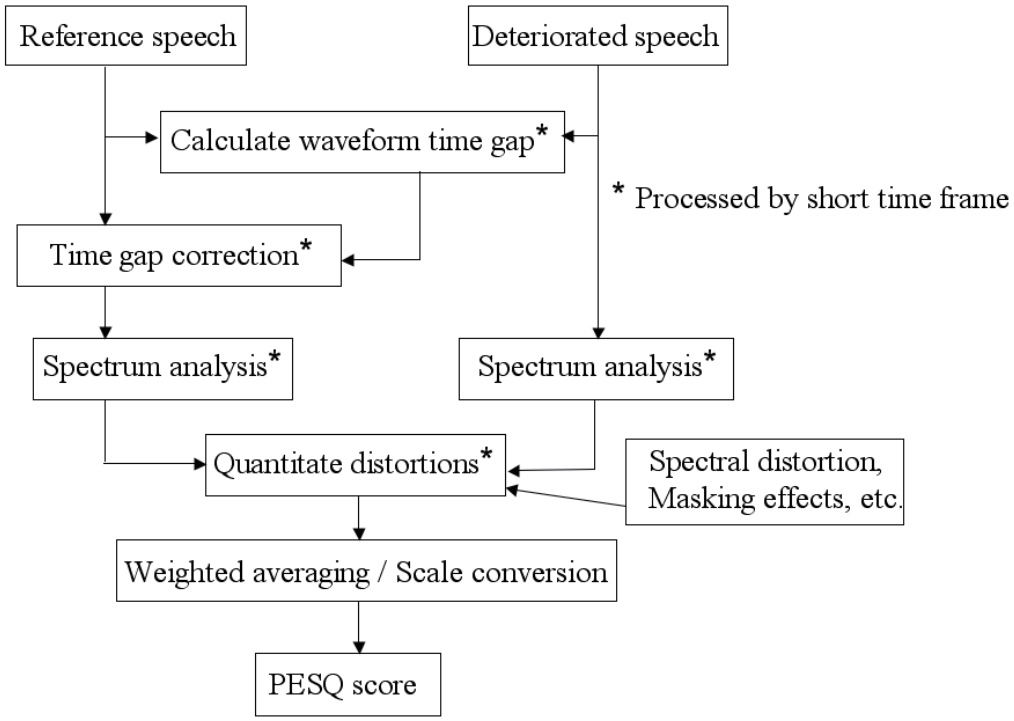

図 4.4: PESQ 値の算出法

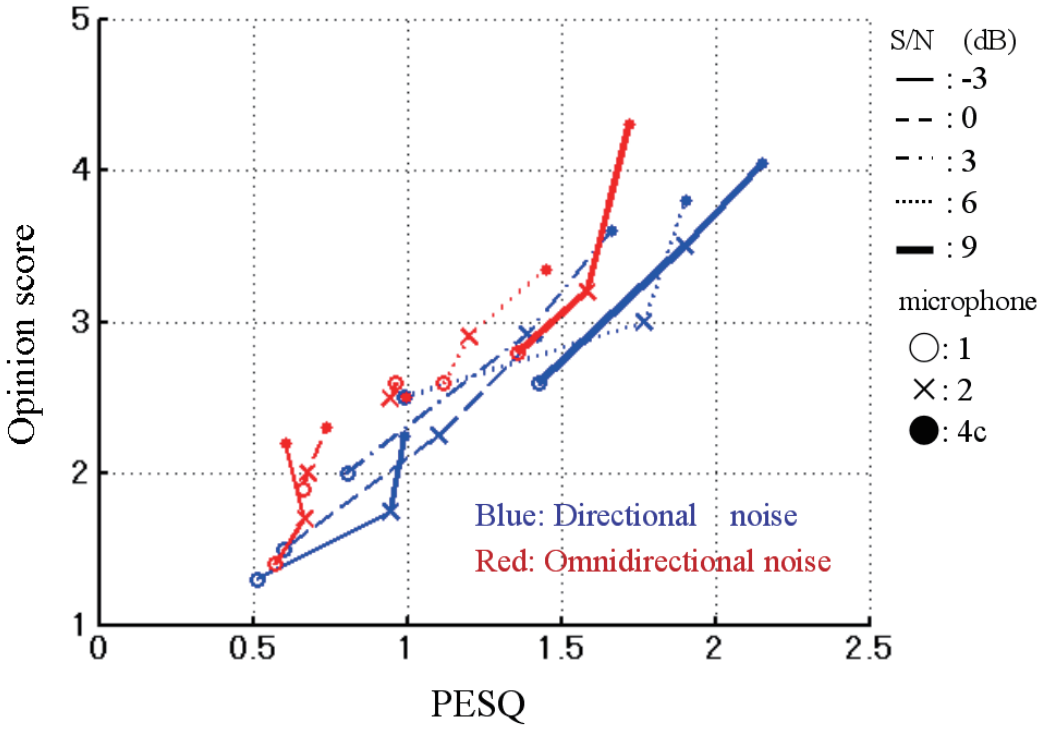

図 4.5: Opinion 評価値と PESQ

評価値と相関が高いことが解る。本研究では、PESQ 値に対し狭帯域包絡線を用いた品質評価予測を検討した。図 4.6 に Opinion 評価値と狭帯域包絡線振幅ヒストグラムの関係を示す。図に示すようにオピニオン評価が良い信号では原信号の小さな振幅も狭帯域包絡線が保存している。一方、オピニオン評価が悪い信号では狭帯域包絡線において原信号の小さな振幅情報は失われ、狭帯域包絡線振幅ヒストグラムが原信号に対し差異が生じることが解る。このことより、Opinion 評価値の予測に振幅ヒストグラム歪みを用いることを考えた。振幅ヒストグラム歪みは原信号に対する収音信号の狭帯域包絡線ヒストグラムの誤差を dB にて表したものである。図 4.7 に Opinion 評価値と狭帯域包絡線振

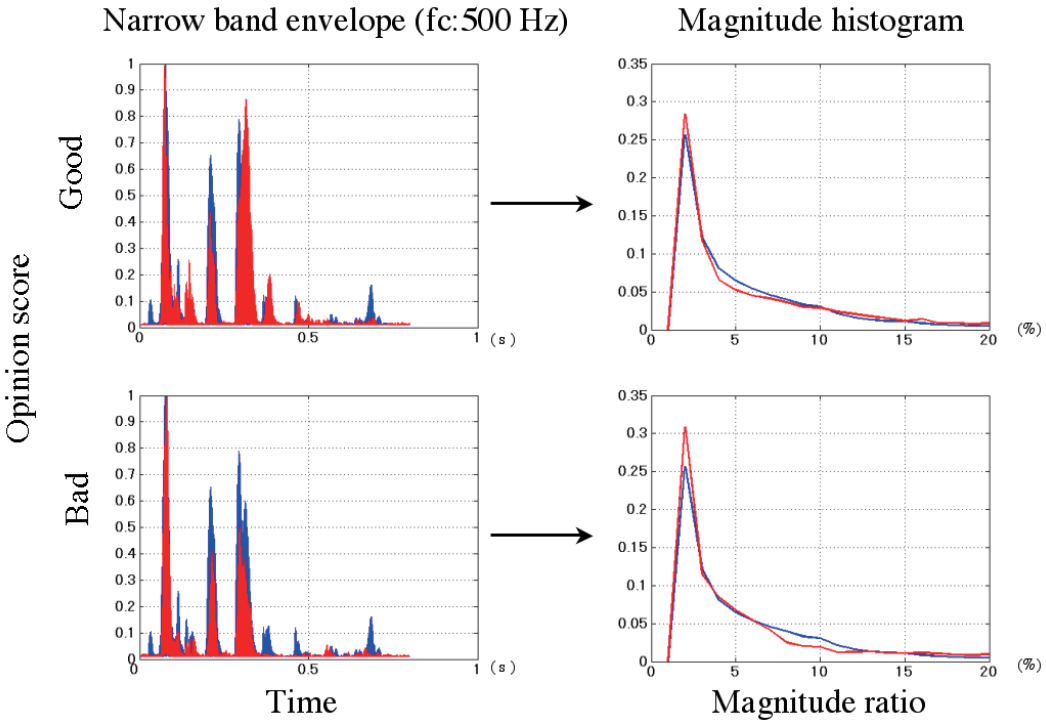

図 4.6: 狭帯域包絡線振幅ヒストグラム (青: 原音声　赤: 収音信号)

幅ヒストグラム歪みを示す。図において、各マイクロホンの数と S/N の向上に比例し Opinion 評価値が上がることがわかる。さらに図より、狭帯域包絡線の振幅ヒストグラム歪みと Opinion 評価値に相関を確認することができる。こ

れにより、狭帯域包絡線が音声明瞭度や音声了解性だけでなく音声信号の品質にも関わる重要な要素を含むことがわかった。次に実環境における収音信号を用いた音声信号の品質評価を行う。

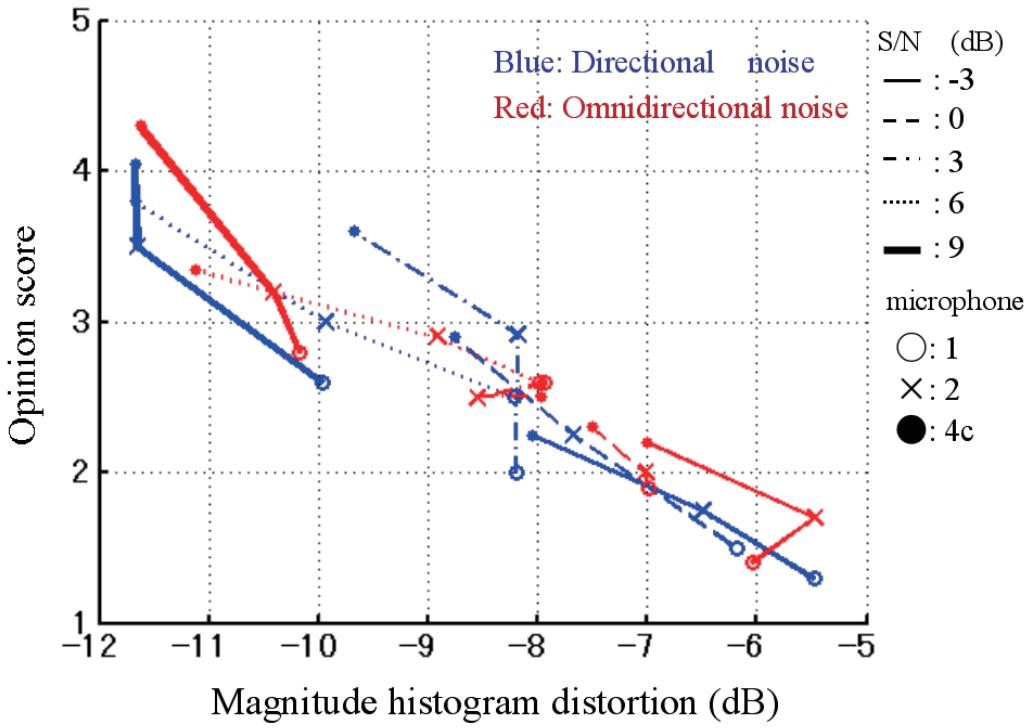

図 4.7: Opinion 評価値と狭帯域包絡線振幅ヒストグラム歪み

4.4　環境騒音下におけるの音声信号品質評価

本節では実際の環境における音声信号を複数収音し Opinion 評価試験を行った。図 4.8 に Opinion 評価結果と従来法である品質評価手法 PESQ との比較結果を示す。図より概ね PESQ により音声品質が評価可能であるが Opinion 評価値が低い場合、PESQ による評価が難しいことが解る。図 4.9 に Opinion 評価値と狭帯域包絡線振幅ヒストグラム歪みの関係を示す。 図より包絡線振幅ヒストグラム歪みによって主観評価値を予測することが可能であることがわかる。

また、狭帯域包絡線の振幅ヒストグラム歪みによる予測では PESQ に比べて評価値が Opinion 評価値が低い場合においても評価可能であることがわかる。

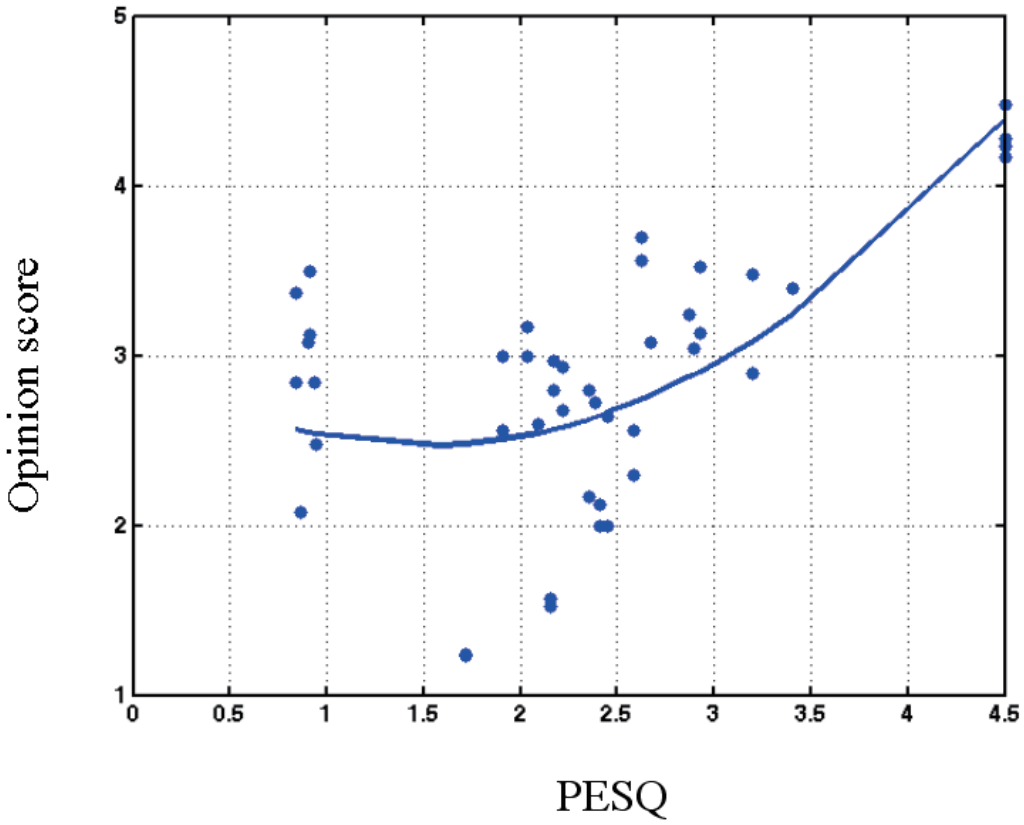

図 4.8: 実環境における Opinion 評価値と PESQ

4.5 考察

本章では位相情報と関連する狭帯域包絡線による Opinion 評価値の予測を試みた。その結果、狭帯域包絡線の再現性が高い収音信号が Opinion 評価値も高い結果となった。これより狭帯域包絡線の振幅ヒストグラム歪みを用いて Opinion 評価値が予測可能であることを示した。さらに、狭帯域包絡線相関係数は従来法である PESQ を用いた予測とも比較した。その結果、Opinion 評価

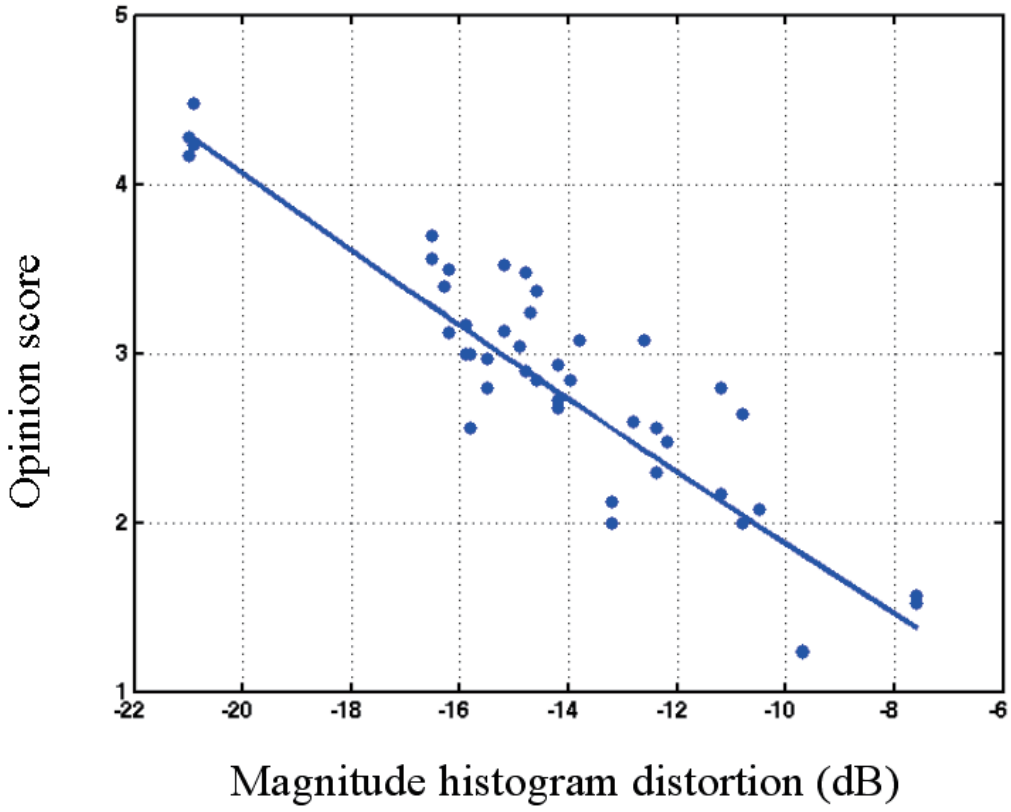

図 4.9: 実環境における Opinion 評価値と狭帯域包絡線振幅ヒストグラム歪み

値と PESQ は実環境下における示したが狭帯域包絡線相関係数は二極化する傾向が少ないことを示した。

4.6 むすび

本章では、音声明瞭度ならびに音声了解度が高い場合の音声信号評価法である Opinion 評価値を用いて音声品質評価実験を行った。また、Opinion 評価値の予測には狭帯域包絡線相関係数を用い包絡線変化と音声品質の関係を検討した。その結果、Opinion 評価値と狭帯域包絡線相関係数に相関があることを示した。これにより、音声了解度の低い場合にも評価可能な音声明瞭度と了解度が高い場合に用いる音質評価の双方において狭帯域包絡線が重要なことを明らかにした。

第5章　包絡線ヒストグラム距離と音声マスキング効果

5.1　まえがき

前章において、無相関な定常雑音と音声信号における音声明瞭度ならびに品質評価を位相情報および狭帯域包絡線を用いて行った。本章では音声信号と音声信号に相関のある音響信号を重畳した信号における音声了解度評価について考察する。相関のある信号が互いに信号情報の聞き分けを困難にする現象としてマスキング効果が知られている。マスキング効果はしばしば公共スペースのプライバシー保護に用いられ [Wang and Bradley, 2002]、エネルギーマスキングと情報マスキングの２つが知られている。エネルギーマスキングによるマスキング効果はSN比と大きく関係する。一方、情報マスキングは聴者の知覚処理において、マスカーの干渉によりターゲット信号の分離と検出を困難にする [Freyman et al., 1999], [Arbogast et al., 2002], [Schmitz and Iyer, 2003]。このように知覚に関わる情報マスキングはエネルギーマスキングより音声のマスキングに効果的だと考えられる。本章では、互いに相関のある音響信号を重畳した信号における音声了解度の検討を行う。

マスキング効果は聴覚システムと非常に関係の深い効果であり、一般的にエキサイテーションパターンに関係する周波数領域のSN比によって評価される [Espinoza-Varas and Cherukuri, 1995]。SN比はエネルギーマスキングの評価に適している。しかしながら、逆再生音声としてよく知られる効果的な情報

マスカー [Rhebergen et al., 2005] のように、ターゲット音声と似たような周波数特徴をマスカーが持っている場合、SN 比が 0 dB であってもそれらの音声の分離は非常に困難となり、単純な SN 比からマスキングを評価することは難しい。このように、聴覚システムにおけるマスキング効果のメカニズムにはスペクトラム情報だけでなく時間波形の時変情報が関係していると思われる。情報マスキングの研究において信号の類似が情報マスキングと深く関係していることが知られている [Hoen et al., 2007] , [Navarro and Pimentel, 2007] 。例えば Durlach 等 [Durlach et al., 2003] はターゲットとマスカーの類似度の低下がマスキング効果を減少させる傾向があることを報告している。

したがって、本章ではマスキング効果の評価に狭帯域包絡線の振幅分布に着目した信号類似度を考えた。信号類似度は包絡線の振幅ヒストグラムのケプストラムより算出するヒストグラムのケプストラム距離（HCD:magnitude histogram cepstrum distance）として提案する。まず、ターゲットとマスカーより包絡線の振幅ヒストグラムのケプストラムを算出し、もしこられのケプストラムがターゲットとマスカーの間で異なれば、その信号は時間領域において類似していないと考える。本章では、時間領域において信号が似ていなければターゲットとマスカーの分離は聴覚上簡単であると考え、HCD により情報マスキングと音声了解度の関係を検討する。情報マスキング効果を伴うマスカーに、音声・音楽・音声の振幅周波数特性をもつ雑音を用いた。音源情報の異なるマスカーを用いた音声了解度評価結果と狭帯域包絡線による情報マスキング評価に関係があるならば、狭帯域包絡線が音声だけでなく音響信号における情報知覚に重要であることが示唆されるものであろう。

5.2 狭帯域包絡線による信号類似度評価

信号類似度は聴覚システムにおける受聴信号の分離と深く関わっている。ターゲット信号とマスカーが大きく異なる場合、二つの信号の検出と分離は容易になる。したがって、本研究では包絡線の時間変化による信号類似度評価を行った。

5.2.1 包絡線ヒストグラムと信号間距離

マスキング効果は一般的にエキサイテーションパターンに関係する周波数領域の SN 比によって評価される [Espinoza-Varas and Cherukuri, 1995]。一方、逆再生音声が情報マスキングのマスカーとしてよく知られている [Rhebergen et al., 2005]。逆再生音声はターゲット音声が全く同じ振幅スペクトル情報を持つ SN 比 0dB の信号である。そして、この信号はターゲット音声と同じ振幅スペクトル情報を持つ雑音と比べ音声了解度が情報マスキング効果により著しく低下する。このことから、情報マスキングが単純な SN 比により評価することが難しいことがわかる。さらに、逆再生音声ではターゲットの知覚も不確かな認識となる。これは合成信号の分離と理解のメカニズムが周波数領域における SN 比だけでなく時間情報と大きく関わっていることを示唆している。したがって、情報マスキングの評価に時間情報を含み音声了解性に重要である狭帯域包絡線を用いる。

Nakashima 等は合成信号の確率密度関数をケプストラムの逆畳み込みにより分離する方法を提案した [Nakashima et al., 1996]。ターゲット信号 X が無相関のマスカー Y と合成されている場合、合成信号 Z の確率密度関数のような自乗包絡線ヒストグラムは以下のように計算できる。

$$p(Z) = p(X) * p(Y), \tag{5.1}$$

ここで $p(*)$ はそれぞれの確率密度関数を表す。確率密度関数のフーリエ変換は特性関数として知られている。従って、確率密度関数の畳み込みは

$$F[p(Z)] = F[p(X)]F[p(Y)] \tag{5.2}$$

$F[*]$: はフーリエ変換を表す。ケプストラムが特性関数から算出可能であれば、確率密度関数はターゲットとマスカーの特性関数に分割できる。よって、ターゲットとマスカーの確率密度関数のケプストラム Cp（Z）は

$$\begin{aligned} Cp(Z) &= F^{-1}[lnF[p(X)]] + F^{-1}[lnF[p(Y)]] &(5.3)\\ &= Cp(X) + Cp(Y). &(5.4)\end{aligned}$$

ここで、ターゲットとマスカーの分割された確率密度関数がケプストラム領域で得られる。また、これらの確率密度関数のケプストラム距離は

$$Dc = \sqrt{\frac{1}{N}\sum (Cp(X) - Cp(Y))^2}, \tag{5.5}$$

ここで、Dc は確率密度関数のケプストラム距離を表す。ヒストグラムケプストラム距離（HCD:magnitude histogram cepstrum distance）は確率密度ケプストラム距離 Dc のように $p(X)$ をターゲット、また $P(Y)$ をマスカーの自乗包絡線の振幅ヒストグラムとし信号の類似度を算出する。したがって、HCD が増加する場合、信号類似度とマスキング効果は減少すると考えられる。

5.2.2　狭帯域包絡線を用いた信号間距離評価

前節では HCD によって信号類似度が評価できると考えた。したがって、音声信号をターゲットとしマスカーと比較する信号類似度実験を行った。ターゲット信号は男声とし信号長２秒の発話音声を用いた。この実験では異なる５つのマスカーを使用した。それぞれマスカーはターゲットと同じ振幅スペクト

ルを持つ雑音・音楽・女声信号・ターゲットではない話者が発話した男声信号・ターゲットと同じ話者が別発話した音声信号をもちいた。

図 5.1 に１/４オクターブバンド、中心周波数 500Hz におけるターゲットとマスカーの自乗包絡線を示す。横軸は時間を縦軸は振幅を表す。パネル（a）ターゲット音声信号の自乗包絡線パネル（b）ターゲット音声の振幅スペクトルを持つ雑音の自乗包絡線パネル（c）音楽の自乗包絡線パネル（d）女声の自乗包絡線パネル（e）ターゲットと異なる話者の男声自乗包絡線パネル（f）ターゲット話者による別発話信号の自乗包絡線自乗包絡線は振幅情報の時間変化を含んでいる。この図から狭帯域包絡線そのものから信号類似度を評価することが難しいことがわかる。まず、定常雑音の包絡線の形は一定で音楽の包絡線はテンポ・リズム・メロディーに依存する。そして、音声信号の包絡線は音声信号の発話内容に依存する。例えば、図 5.1a において (f) は同一話者の同一文章にも関わらず発話が異なるだけでピーク位置も異なることがわかる。

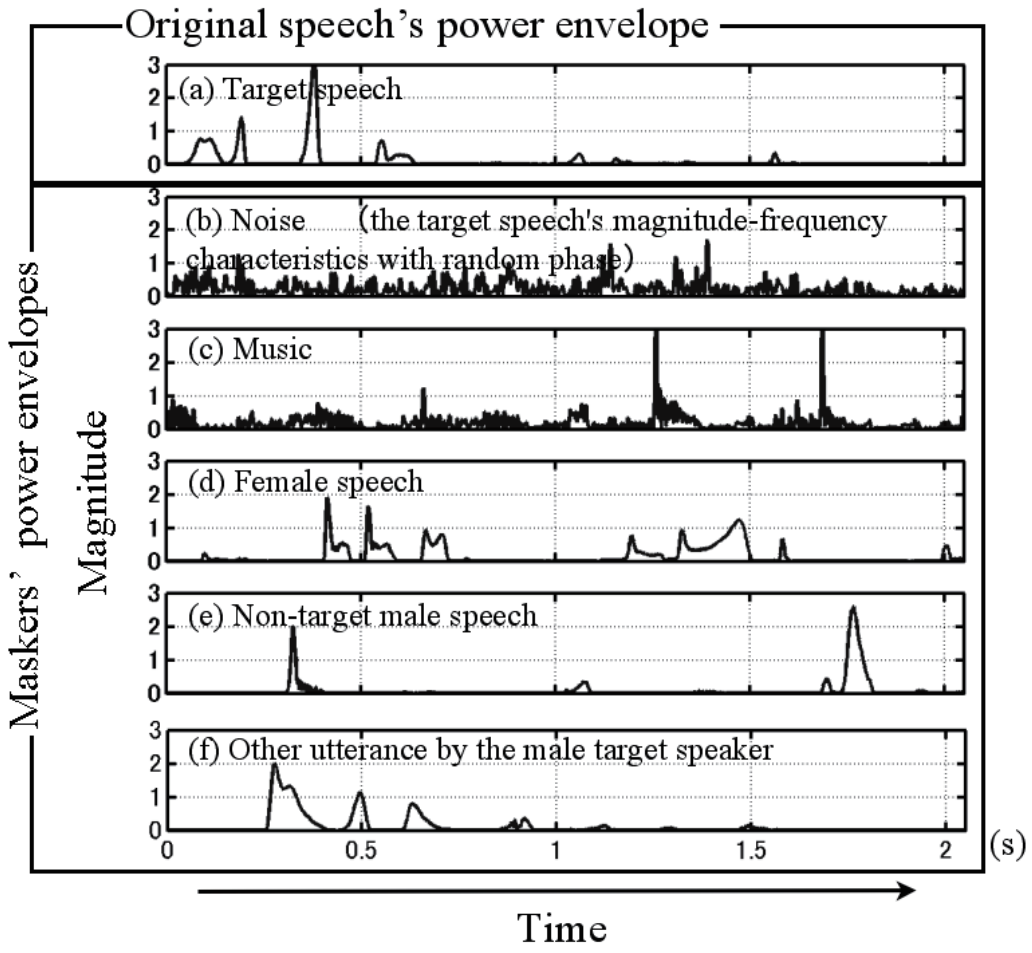

図 5.1: ターゲットとマスカーの自乗狭帯域包絡線 (fc: 500 Hz)

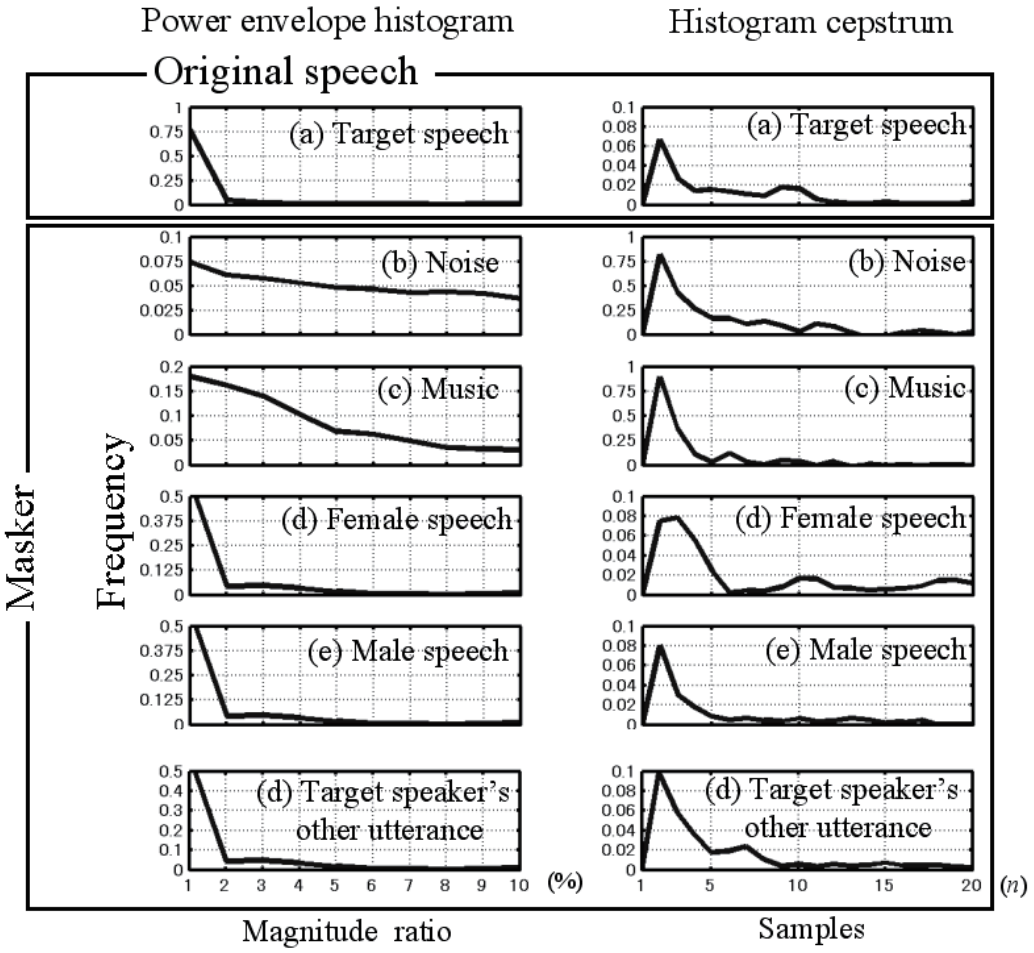

図 5.2: 包絡線振幅ヒストグラムとケプストラム (fc: 500 Hz)

図5.2は１/４オクターブバンド、中心周波数５００Hzにおける自乗包絡線ヒストグラムとそれぞれのケプストラムを示す。パネル（a）ターゲット音声信号パネル（b）ターゲット音声の振幅スペクトルを持つ雑音パネル（c）音楽パネル（d）女声信号パネル（e）ターゲットと異なる話者の男声信号パネル（f）ターゲット話者による別発話信号左のパネルは自乗包絡線のヒストグラムを示す。横軸は自乗包絡線の振幅最大値を１００％とする振幅率とし、縦軸は包絡線振幅の分布を表す。自乗包絡線のヒストグラムにおいてもそれぞれの信号は似た傾向を示している。右のパネルはヒストグラムのケプストラムを示す。この結果、ヒストグラムのケプストラムにおける最大値は雑音が最も大きくなり、さらに音楽信号が音声信号に対し大きくなることが確認できる。このようにヒストグラムのケプストラムにより信号間に差があらわれた。さらに、音声信号はそれぞれ似た傾向を持つことがわかる。このことから、HCDによる信号類似度評価が可能であると考え分析を行った。

図 5.3 に男声音声信号をターゲットとし比較した信号類似度を示す。横軸は１/４オクターブバンドにおける中心周波数を表し、縦軸は HCD を表す。この図において、音楽と雑音は音声信号と大きな距離を持つこと示した。これは HCD による信号類似度評価の信頼性を示すものである。図 5.4 は１/４オク

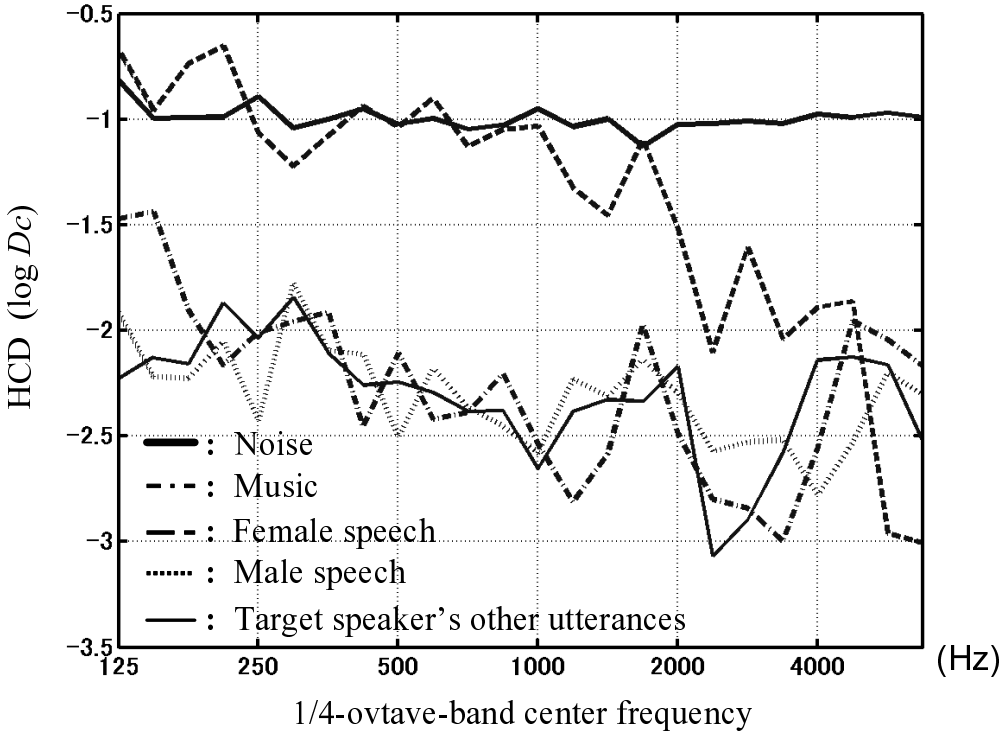

図 5.3: 1/4 オクターブバンドにおける HCD

ターブバンド毎の距離を全ての帯域にわたり平均した HCD を示す。この結果、音声の HCD は約-2.4 から-2.3 となった。また、図より音楽・雑音と音声の HCD を比べると互いに類似度が低いことが容易に確認できる。ここで、HCD による信号類似度評価を議論した。その結果、HCD による信号類似度と情報マスキングの関係を考えた。次に HCD を用いたマスキング効果の評価を行った。

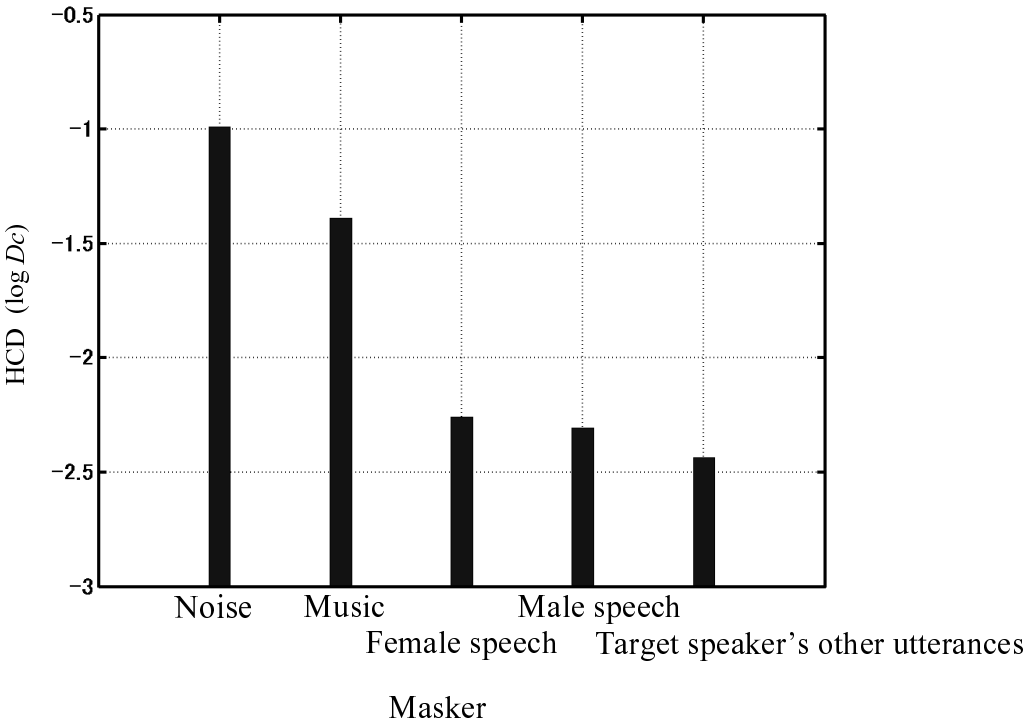

図 5.4: マスカー信号と HCD

5.2.3 バブルスピーチによる信号間距離評価

バブルスピーチは複数の音声信号を用いて作られる音声の情報マスカーとして知られている [Van Engen and Bradlow, 2007]。多くの音声がバブルスピーチの作成に用いられる場合、情報マスキングは低下する [Drullman and Bronkhorst, 2004]。これは複数音声の妨害によりマスカー自身の音声情報も低下することに起因する。本節ではバブルスピーチを用いて HCD による信号類似度と情報マスキングの関係を考えた。

図 5.5 にバブルスピーチの HCD を示す。横軸はバブルスピーチ作成に用いた発話数、縦軸は 1 / 4 オクターブバンドの帯域平均 HCD を示す。実線はターゲット音声話者による別文章発話を用いて作成したバブルスピーチを、点線はターゲット音声と異なる話者の発話による音声信号を用いたバブルスピーチを示す。この図からバブルスピーチ作成に用いる発話数の増加に伴い HCD も大きくなることがわかる。また、同一話者によるバブルスピーチの場合、別話者によるバブルスピーチと比べ HCD 距離が小さくなることわかる。しかし、7 つ以上の発話を合成した場合、同一話者と別話者によるバブルスピーチが同様の傾向を示す結果となった。

これより、HCD により情報マスキングが評価可能であることを確認した。しかし、HCD と聴覚における情報マスキングの関係はいまだ不鮮明である。次に音声認識関わる音声了解度と HCD の関係を考察する試聴実験を行った。

5.3 会話完成率に着目した音声マスキング評価実験

2 つの信号から逆再生信号の特徴を保存するダブルマスカーを作成し情報マスキング評価実験を行った。

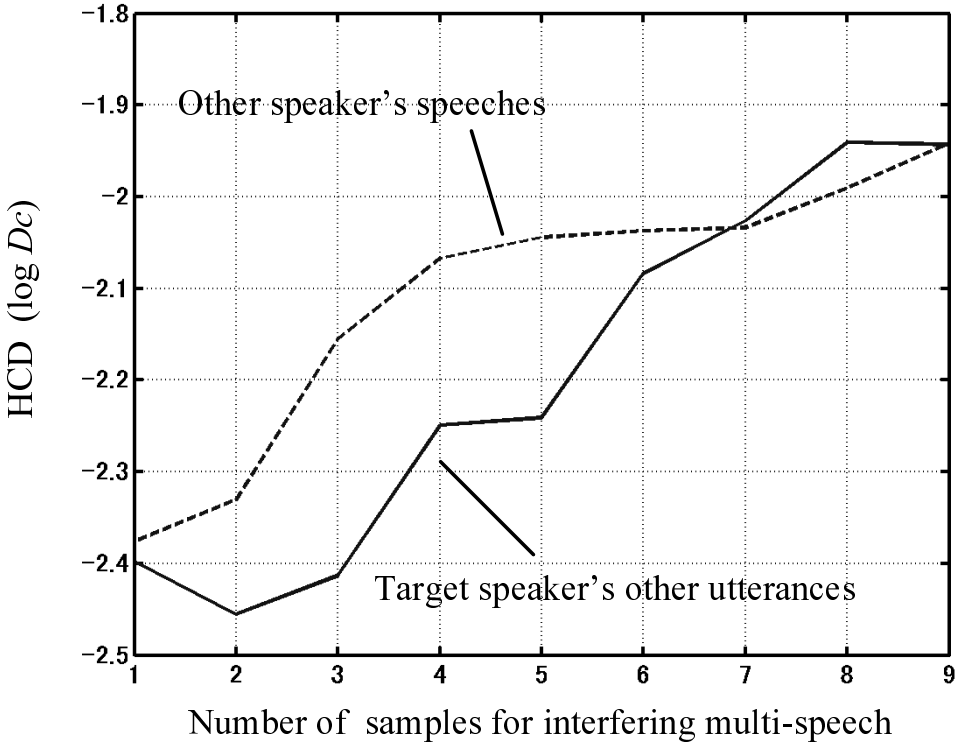

図 5.5: バブルスピーチと HCD

5.3.1 逆再生音声信号の特徴

逆再生音声は効果的な情報マスカーとして知られている [Rhebergen et al., 2005]。ターゲット音声がターゲット音声の逆再生信号と混ざった場合、ターゲット音声の分離と理解は困難になる。本実験では、この逆再生音声の特徴を考慮し効果的なマスカーを作成した。原音声信号スペクトルを $X(k)$ とすると逆再生音声信号スペクトラムは $\hat{X}(k)$ となり、

$$\hat{X}(k) = X^*(k), \tag{5.6}$$

となる。ここで、k は周波数番号を示す。このように逆再生音声スペクトラムは音声信号の複素共役となる。したがって、原音声と逆再生信号の合成信号は

$$X(k) + X^*(k) = 2 \times real[X(k)] \tag{5.7}$$

となる。合成信号は実部スペクトル情報のみ保有し信号長の半分から折り返す信号となる。次に HCD と音声認識の関係を明らかにするためにダブルマスカーを作成する。

5.3.2 ダブルマスカーの作成

試聴実験によって情報マスキングと HCD の関係を明らかにするためにダブルマスカーを作成する。図 5.6 にダブルマスカー作成の概要を示す。ダブルマスカーの作成には２つの音源を必要とする。まず、信号 A と B 両方をフーリエ変換により実部と虚部のスペクトラムに分割する。そして、虚部を互いに－１倍し信号間で入れ替える。このようにして再合成した信号を両信号の特徴を保有しているダブルマスカーとした。試聴実験には４種類のダブルマスカーを作成した。それぞれのダブルマスカーは音声と雑音・異なる２音声・音楽と音声・ターゲット音声を含む異なる２音声とした。この信号を用いた音声了解度

は試聴実験を通して評価した。また、ダブルマスカー作成には無音区間のない音声信号を、音楽は作成したものを使用した。

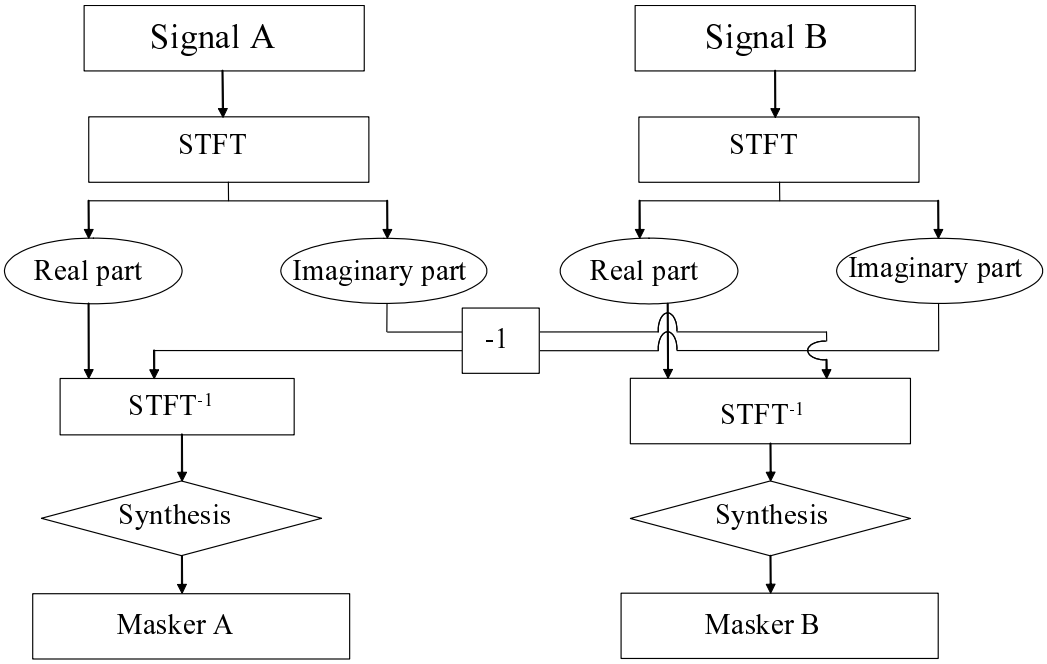

図 5.6: ダブルマスカー作成方法

5.3.3 会話完成率を用いたマスキング効果評価

情報マスキングが生じる合成信号からターゲット信号が検出または分離可能であるかを検証するためにしばしば”不明確さ”が検討される [Durlach et al., 2003]。したがって、本実験では情報マスキングが起こっている音声の不明確さを測るために会話完成率を考案した。書き取りによる音声明瞭度は一般的にマスキング効果の評価に用いられる。しかしながら、会話中の内容はしばしば音声明瞭度が低い場合においても理解することができる。したがって、音声明瞭度による情報マスキングの評価は難しい。しかし、会話完成率は答えと比較する必要のない質問を作成し、回答により会話が成り立つ場合に正答とし判断

する。ターゲット音声が情報マスキングの増加により不明確な場合、被験者は質問を聴き分けることが困難となり適正に質問に回答することができない。したがって、マスキング効果があるとき会話完成率は下がることとなる。

5.3.4 マスキング効果評価結果

この試聴実験は作成した４つのダブルマスカーを用いて行った。ダブルマスカーに用いた信号は音声とそれぞれ (a) 音声振幅スペクトルをもつ雑音によるマスカー、(b) 音楽マスカー、(c) 音声マスカー、(d) ターゲット信号の虚部を用いた音声マスカーとした。実験にあたり１２０個の質問を無響室で収音し、質問は試聴実験の試験音としてマスカーと合成し使用した。全ての質問は簡単に答えられる個人的な質問とし、そのほとんどは１単語で回答できる質問である。試験音数は１２０サンプル、各条件１０サンプルづつ、SN 比の条件は－６、－３、０ dB の３条件とした。被験者は２２～２７歳の７名の男性で日本語を母国語とする健聴者とし、試験信号をヘッドホン (AKG-K240) から好みの音量でダイオティックコンディションで試聴した。また、被験者には”はい”と”いいえ”の回答を禁じた。

図 5.7 に試聴実験結果を示す。横軸は音声とのダブルマスカー作成信号を示し、縦軸は会話完成率を示した。実線はターゲットとマスカーの SN 比－６ dB、破線は－３ dB、一点鎖線は SN 比０ dB の試験結果を示す。本実験では、試聴実験結果に分散分析（ANOVA）を行った。一元配置分散分析は４つの信号種類と 3 SN 比条件に対して実施した。その結果、”信号の種類”の差は有意である結果となった ($F(3,8) = 54.31$,$p < 0.001$)。一方、SN 比条件は有意でない結果となった ($F(2,9) = 0.14$,$p > 0.5$)。この結果から、信号種類が SN 比に比べ大きく効果的であることがわかる。また、試験結果からターゲット音声を用いたダブルマスカーでは音声の内容がほぼ完全に理解できないことが確認でき

る。さらに、音声信号を用いたマスカーは音楽・雑音マスカーに比べ大きくマスキング効果が現れることがわかる。

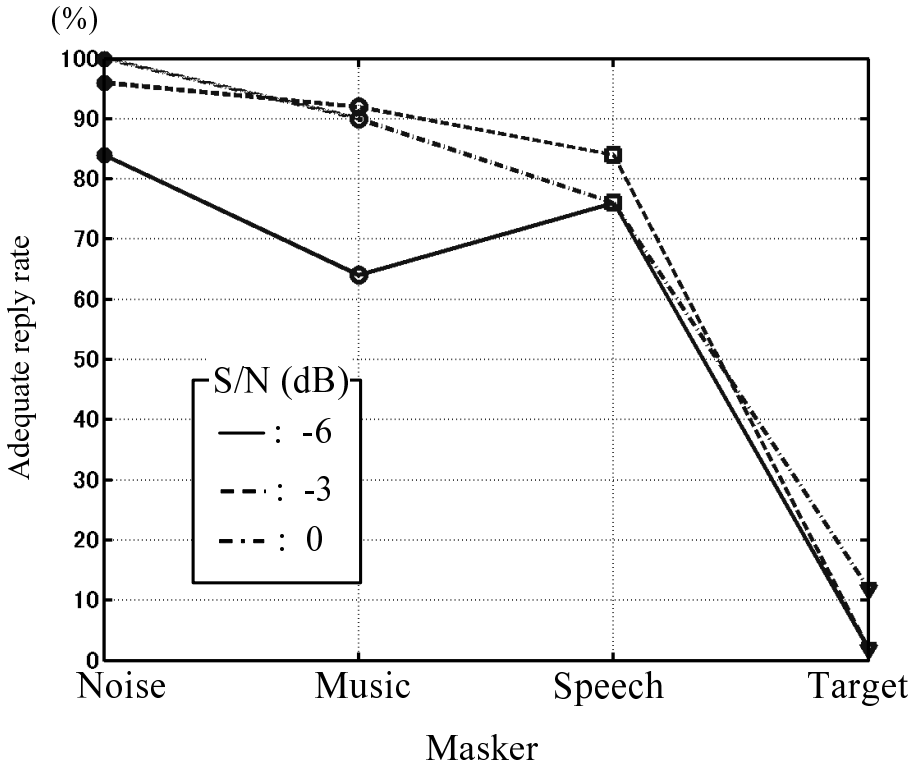

図 5.7: 会話完成率による試聴実験結果

図 5.8 はダブルマスカーとの HCD による信号類似度を示す。横軸は音声とのダブルマスカー作成信号を示し、縦軸は HCD を示す。HCD の信号類似度は SN 比に依存しない。したがって、本実験では SN 比別の試験信号で算出した HCD を平均した。この結果より、雑音を用いたマスカーが４信号の中で最もターゲットと離れていることがわかる。また、音楽を用いたマスカーは音声を用いたマスカーより離れ、ターゲット音声を用いたマスカーの距離が最も近い結果となった。これらの結果より、HCD は試聴実験結果と同様な傾向を持つことがわかった。この結果は HCD により情報マスキングの評価が可能である

ことを示している。帯域別 HCD では、中心周波数 500Hz において試聴実験結果と一致している。また、中心周波数２ kHz では雑音と音楽を音声と比較した場合中心周波数５００ Hz の HCD と比較し距離が大きいことがわかる。この結果は全狭帯域 HCD を平均することにより安定した結果が得られることを示している。これにより、情報マスキングと信号類似度の関係を改めて確認した。

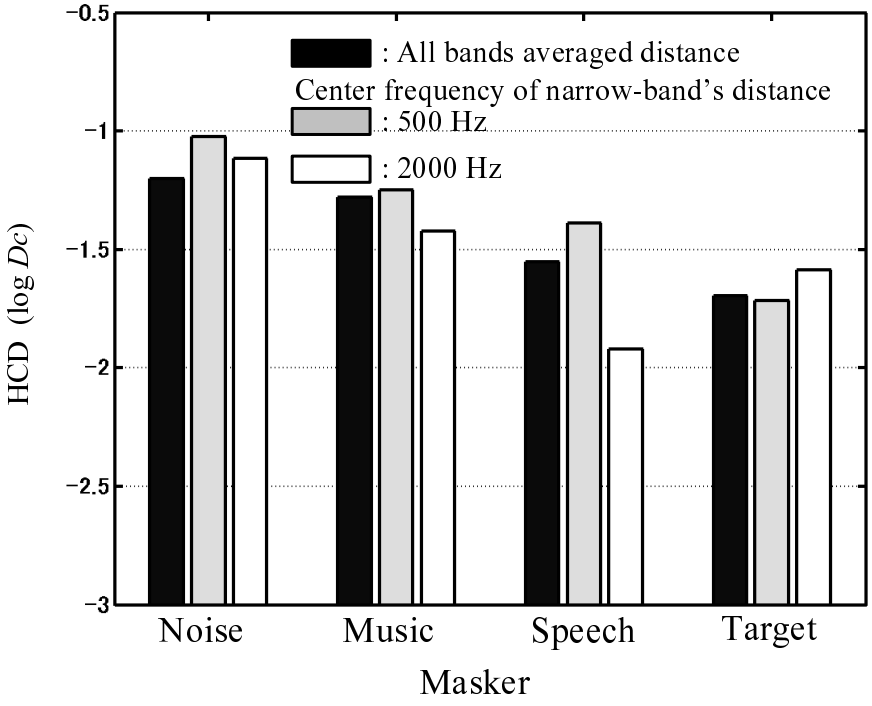

図 5.8: ターゲット音声とダブルマスカーによる HCD

5.3.5 むすび

本章では情報マスキングのための信号類似度を検討した。類似信号が混ざった場合、聴覚において音源情報の理解と分離が困難になる。この情報マスキングに関わるとされる信号類似度を時間波形包絡線の振幅ヒストグラムのケプストラム距離（HCD:magnitude histogram cepstrum distance）により計算し

た。この HCD による信号類似度はバブルスピーチを用い検証した。多くの発話をバブルスピーチに用いた場合、HCD が大きくなることから情報マスキングと同一の傾向を持ち、HCD が情報マスキング効果と対応していることを示した。さらに、HCD による情報マスキング評価を行うために試聴実験を行った。試験音として逆再生音声と同じ特徴を持つダブルマスカーを作成し評価した結果、ターゲット音声を用いたダブルマスカーが最も情報マスキング効果を持つ結果となった。また、音声を用いたダブルマスカーは雑音や音楽に比べ大きな情報マスキングを持つことがわかり、HCD も試聴実験と同様な傾向を示す結果となった。この結果、音源情報の異なるマスカーによる情報マスキングも狭帯域包絡線により評価可能であり、狭帯域包絡線が音声だけでなく音響信号における情報知覚に対しても重要であることを示した。

第6章　狭帯域包絡線に着目した話者知覚

6.1　まえがき

前章において、互いに相関のある信号を用いた音声了解度を狭帯域包絡線に着目し評価した。さらに、種類の異なる音響信号を用い狭帯域包絡線に基づく信号類似度評価を行った結果、狭帯域包絡線に音源情報が含まれることを示した。本章では、音源情報として話者情報に着目し狭帯域包絡線との関係を試聴実験により明らかにする。

信号波形における狭帯域包絡線は音声情報を得る大きな手掛かりとして知られている。Drullman 等 [Drullman, 1995] は 100Hz-6.4kHz にわたる 24 帯域の 1/4 オクターブ帯域包絡線とそれぞれの周波数帯域に対応する帯域雑音から了解性のある音声を合成できること、Shannon 等 [Shannon et al., 1995] は音声帯域を概ね 4 帯域に分割した帯域包絡線を用いて了解性のある音声が実現できることを示した。また Houtgast 等 [Houtgast and Steeneken, 1973] [Houtgast and Steeneken, 1985] は狭帯域包絡線の変調度によって、室内の反射音や雑音による了解性劣化の評価が可能であることを報告している。このように狭帯域包絡線が音声の了解性に関わる重要な要因であることが明らかにされてきた。しかしこれらの先行研究ではいずれも話者情報の保存あるいは復元については言及されていない。

一方、Li 等 [Li and Hughes, 1974] はスペクトルの時間変化に着目した話者

特徴分析により狭帯域包絡線に含まれる話者情報に言及した。これに続いて風間等 [風間道子他, 2009] および Gotoh 等 [Gotoh et al., 2006] は狭帯域包絡線の包絡線相関行列による話者判定の可能性を示した。

上記の研究より、狭帯域包絡線が音声情報の知覚と関わっていることから、狭帯域包絡線に含まれる話者情報の知覚に興味を抱いた。そこで狭帯域包絡線に着目した試聴実験による話者判定実験を試みることとした。試験信号は二人の話者の音声信号から狭帯域包絡線を互いに入れ替えることにより合成信号を作成した。試聴による話者判定実験は XAB 法によるものとし、原話者信号を被験者に提示し被験者は続く一組の合成信号において原話者と同一話者と判断した信号を選択する。本章では狭帯域包絡線が、人間が話者を判定する上で有効な話者情報を有していることを明らかにする。

6.2　狭帯域包絡線と搬送波に着目した話者判定試聴実験

信号波形の振幅時間変化をあらわす包絡線、特に狭帯域包絡線は人口蝸牛に用いられている [Wilson et al., 1991]。これに伴い、狭帯域包絡線と微細構造となる搬送波の知覚における関係が研究されている [Smith et al., 2002]。本章では合成信号を用いた話者判定試聴実験を行い、狭帯域包絡線・搬送波と話者知覚の関係を考察する。

6.2.1　試験音作成

包絡線と搬送波における話者特徴を検証するため、二人の異なる話者 A・B による音声信号を用いた合成信号の作成を行った。図 6.1 に信号合成に用いる狭帯域包絡線と搬送波の合成例を示す。図 6.1a,e に異なる話者の発話による音声信号 A,B を中心周波数 125Hz の 1/4 オクターブバンドフィルタに通した狭

帯域波形を示す。図 6.1b,f は信号 A,B の狭帯域信号より算出した包絡線、図 6.1c,g は搬送波をそれぞれ示す。図 6.1d は信号 B の包絡線 (図 6.1f) と信号 A の搬送波 (図 6.1c) を積により合成した合成信号、図 6.1h は信号 A の包絡線 (図 6.1b) と信号 B の搬送波 (図 6.1g) を積により合成した合成信号を示す。このように異なる音声間の狭帯域信号より得られた包絡線と搬送波を入れ替え、積をとることにより再合成し狭帯域合成信号とした。

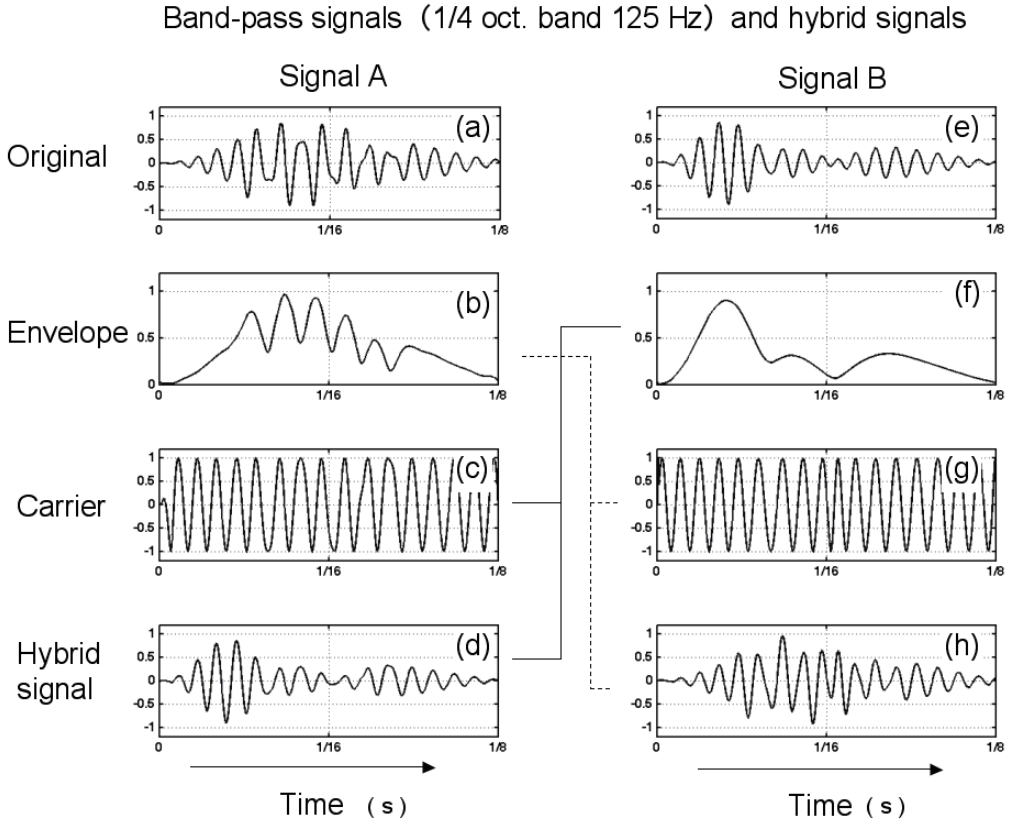

図 6.1: ヒルベルト変換による信号合成例 (a) 信号 A の狭帯域信号, (b) 波形 a の包絡線, (c) 波形 a の搬送波, (d) 信号 B の狭帯域包絡線と信号 A の狭帯域搬送波, (e) 信号 B の狭帯域信号, (f) 波形 b の包絡線, (g) 波形 b の搬送波, (h) 信号 A の狭帯域包絡線と信号 B の狭帯域搬送波

Drullman は狭帯域包絡線を用いた試聴実験において、聴覚の特性から 24 帯

域にわたる 1/4 オクターブバンドを使用した [Drullman, 1995]。また、風間等は狭帯域包絡線の相関行列を用いた話者識別実験において、125〜2,000Hz の 1/4 オクターブバンドフィルタと、第三ホルマントの帯域幅を考えた 2,000Hz 〜11,313Hz の 1/8 オクターブバンドフィルタを同時に使用した [風間道子他, 2009]。本実験は試聴実験であることから、聴覚の特性をふまえ合成信号の作成には周波数範囲 125〜11,313Hz の 27 帯域からなる 1/4 オクターブバンドパスフィルタ (”MATLAB Signal Proccessing Toolbox”の 3 次の Butterworth) を用いた。図 6.2 に示すように 2 つの信号はぞれぞれバンドパスフィルタにより

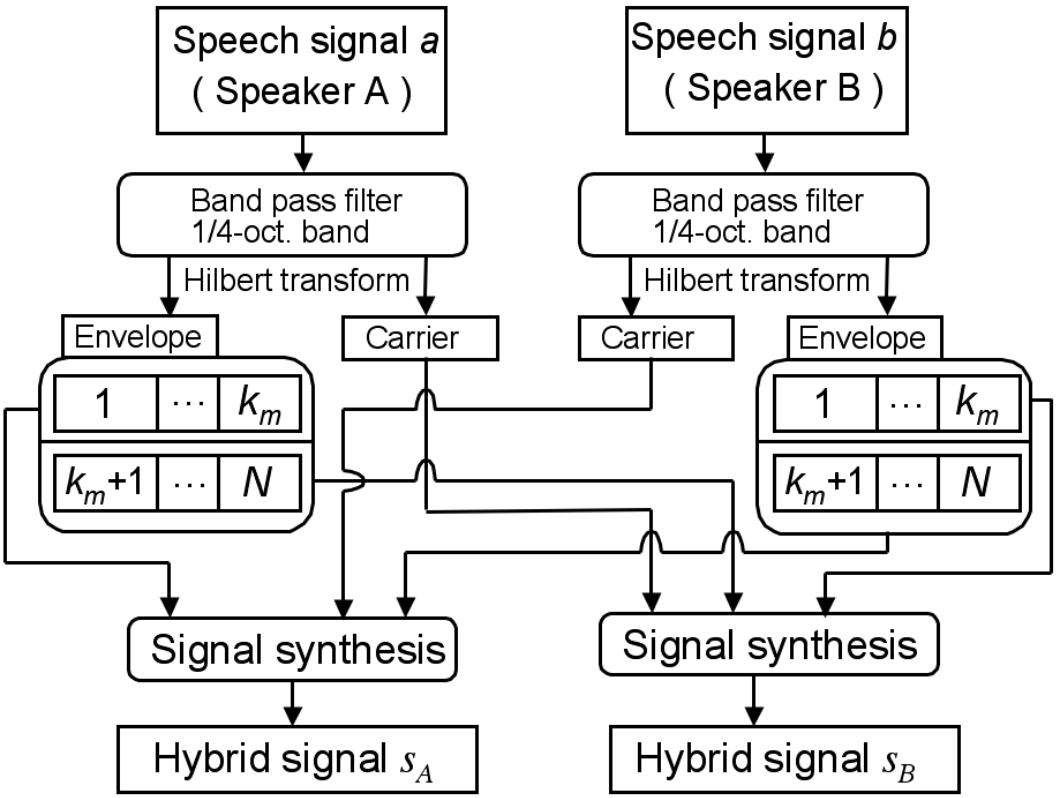

図 6.2: 合成信号作成手順

帯域分割する。分割した帯域信号からヒルベルト変換を用い、狭帯域包絡線と搬送波をそれぞれ抽出する。合成信号は話者 A と話者 B の包絡線を入れ替え

それぞれの帯域の話者 B の搬送波と合成した S_A と、話者 B と話者 A の包絡線を入れ替え話者 A の搬送波と合成した S_B とした。さらに、Gonzalez 等が行った実験は人工蝸牛における狭帯域包絡線と話者特徴知覚を考慮し周波数帯域の分割数を 3,4,5,6,8,10,12,16 と変えたことに対し、本試験では狭帯域包絡線と搬送波の周波数帯域と話者特徴の関係を検証するため low pass type(LP 形) と High pass type(HP 形) の異なる二つの合成信号を作成した。図 2 に示す LP 形の包絡線は最も低い中心周波数 (125Hz) となる帯域番号 1 からカットオフ周波数となる帯域番号 k_m まで話者 A(B) の包絡線を保存し、帯域番号 k_m から最大帯域番号 N における包絡線を話者 B(A) と入れ替えた。ここで、カットオフ周波数となる帯域番号 k_m の中心周波数を包絡線保存周波数とした。また、k_m は LP(あるいは HP) 形における包絡線保存周波数となる帯域番号、N は信号全体の上限帯域番号とした場合、LP 形合成信号における話者の狭帯域包絡線と搬送波の関係は

$$S_{aL} = \sum_{n=1}^{k_m} E_{\mathrm{A}n} C_{\mathrm{B}n} + \sum_{m=k_m+1}^{N} E_{\mathrm{B}m} C_{\mathrm{B}m}, \tag{6.1}$$

となり、HP 形合成信号は

$$S_{aH} = \sum_{n=1}^{k_m} E_{\mathrm{B}n} C_{\mathrm{B}n} + \sum_{m=k_m+1}^{N} E_{\mathrm{A}m} C_{\mathrm{B}m}. \tag{6.2}$$

と表すことができる。ここで $E_{\mathrm{A}*}$,$E_{\mathrm{B}*}$ は話者 A,B それぞれの狭帯域包絡線、同様に $C_{\mathrm{A}*}$,$C_{\mathrm{B}*}$ は狭帯域搬送波を示す。

合成信号作成に用いた音声信号の長さは約 2 秒である。話者は日本人 (男性 25 名、女性 28 名) で 11 の日本語単文をサウンドブースで標本化周波数 48kHz, 量子化ビット 16bit を用いて録音した。実験に用いた発話文を以下に示す。

1. [発話者の氏名] です
2. このじはとおくからみえにくい
3. つもったばかりのゆき

4. きせきがおこった

5. きたかぜがまどにふきつける

6. すぐれたぼうすいせい

7. きょうみぶかいけんきゅう

8. このみちはにしにむかっている

9. かめらのそうさほうほう

10. しゅっせきしゃはなまえをかく

11. はっきりしたひょうげん

ここで文 1 は発話者の氏名 (仮名) を発話する。次節に合成信号と XAB 法を用いた話者判定試聴実験方法を述べる。

6.2.2　XAB 法を用いた話者判定試聴実験

本試験では被験者と試験信号の話者は面識がなく被験者が話者情報を記憶していない事から、原音声による話者提示信号に続いて合成信号を提示する XAB 法を用いた。刺激音の構成並びに提示順序を図 6.3 に示す。本試聴実験

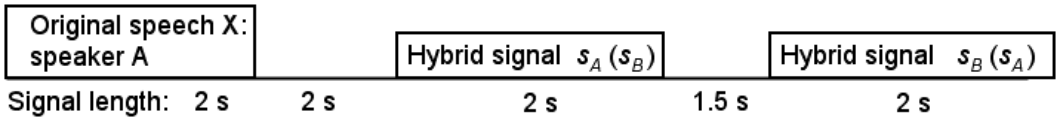

図 6.3: 試聴実験信号の構成

は原信号 X を提示した後、2 秒おいて合成信号 1 と合成信号 2 を 1.5 秒間隔に再生する。回答方法は原信号 X と同一の話者と思われる試験信号 (合成信号 1 あるいは合成信号 2）を選択する。

合成試験信号の組 (S_A,S_B) には LP 形 (S_{A_L},S_{B_L}) と HP 形 (S_{A_H},S_{B_H}) を作成した。実験には 216 サンプル (それぞれの LP 形と HP 形の合成信号において

包絡線保存周波数の異なる 27 通りの刺激音を順序効果を考慮した 8 サンプル) の試験信号を作成した。刺激音作成に用いた音声信号の話者は男性 25 名、女性 28 名から無作為に 2 名を抽出した。ただし男声－女声の組み合わせは無い。原音声には全て文 1 を用い、合成音声の組には文 2〜11 中から無作為に抽出した異なる 2 つの文の組を利用した。被験者は 8 名 (年齢 21 〜 49 才) とし、ヘッドホン (AKG K240) を介して diotic 受聴を行った。被験者は S_A,S_B のどちらが原信号 X に近いかを必要に応じて聞き直しをした後回答した。試験の話者判定は原信号話者 A に対して合成信号話者 S_A と答えた場合を正答とした。本試験において不正答となる合成信号 S_B は原信号話者 A の搬送波を全帯域保有することから、搬送波における話者情報が支配的な場合、全て合成信号 S_B を選び話者判定率は狭帯域包絡線保存周波数と無関係な傾向を示すことが考えられる。

6.2.3 試聴実験結果

図 6.4a,b に男声合成信号試聴実験における LP 形,HP 形の話者判定結果を示す。横軸は包絡線保存周波数、縦軸は話者判定 (正解) 率である。図 6.4a の LP 形における試験結果は包絡線保存周波数が増加すると共に話者判定率が上昇することを示した。また、図 6.4b の HP 形における試験結果は包絡線保存周波数が増加すると共に話者判定率が低下することを示した。LP 形において包絡線保存周波数が上昇すると原話者の狭帯域包絡線を保存する帯域数が増加する。反対に、HP 形は包絡線保存周波数が上昇すると狭帯域包絡線を保存する帯域数が減少する。これより、図 6.4 の結果から包絡線保存帯域数と話者判定率が高い相関関係にあることが読み取れる。

本実験において搬送波が話者判定に優位な場合、不正解となる合成信号が原話者の全ての帯域の搬送波を保有することから、試験結果において全ての包絡

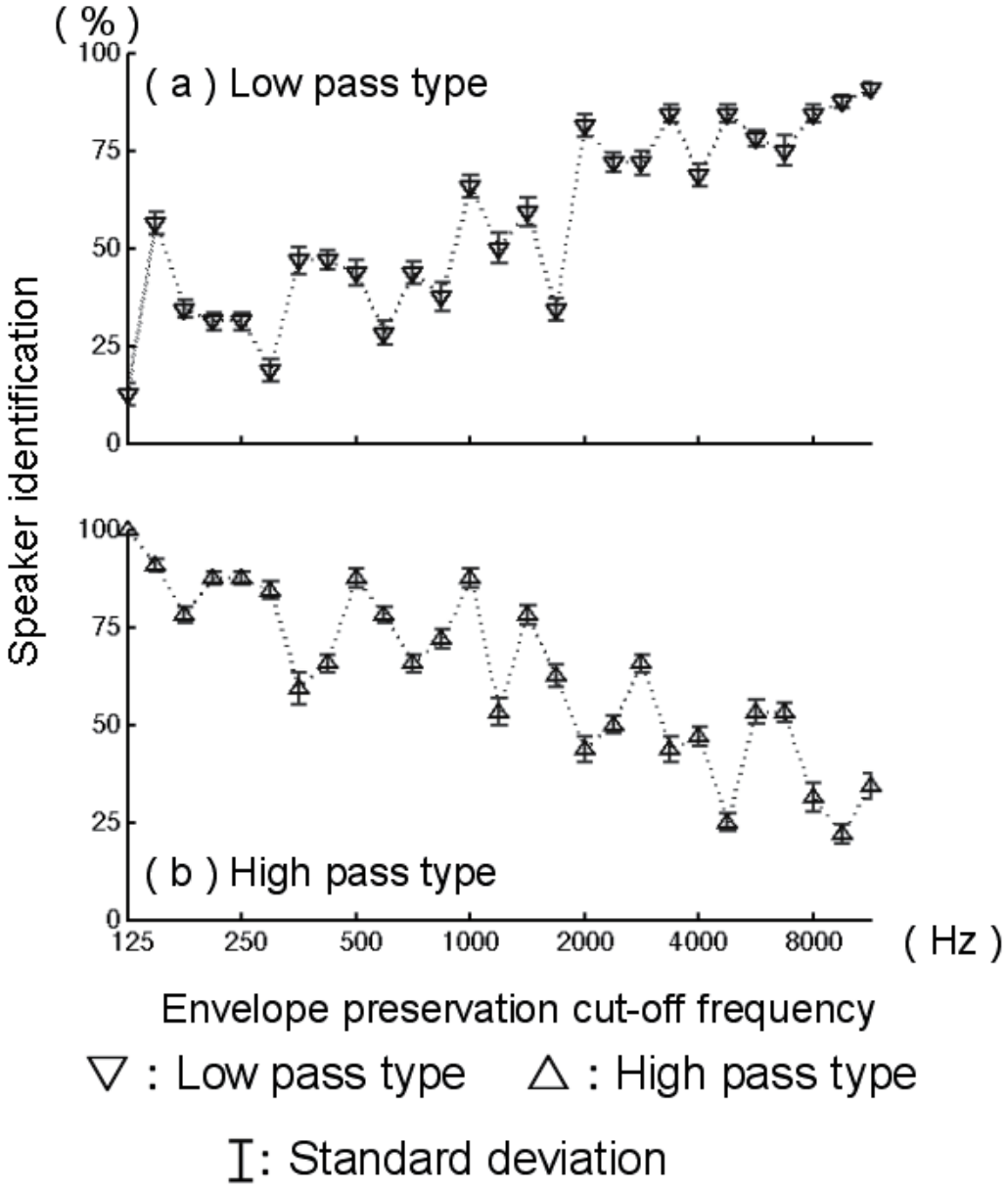

図 6.4: 男声合成信号試聴実験における話者判定率

線保存周波数の話者判定率が一律に著しく低くなると考えられる。しかし、実験結果がその傾向を示していないことから、狭帯域包絡線が搬送波に対し話者知覚の上で優位であることがわかる。

また、LP 形の話者判定率が 75%を上回るには少なくとも低域から少なくとも 2000Hz を含む帯域包絡線が必要であることを示している。一方、HP 形において話者判定率 75%以上を得るには少なくとも 500Hz から高域の情報を併せ持つ必要があることを示している。図 6.5 に女声合成信号に対する試験結果を図 6.4 と同様に示す。図 6.5 から女声の話者判定においても包絡線保存周波数の上昇とともに LP 及び HP 形の話者判定率が上昇及び低下する傾向を示した。このことより、女声における話者判定率も包絡線保存帯域数と高い相関関係を示す結果となった。女声合成信号による試聴実験においても、不正答となる合成信号は原話者信号の狭帯域搬送波を全て保存することから、試験結果より、狭帯域包絡線が搬送波に対し話者知覚に重要である事が考えられる。さらに、女声合成信号においても話者判定率 75%を得るには低域と高域の狭帯域包絡線が共に保存される必要があることが示された。この結果より、連続音声の話者知覚に重要とされる低域のスペクトル情報と、音声生成において個人の声道と関わる梨状窩が影響を与える周波数範囲 2,000〜6000Hz が共に話者知覚に必要である事が考えられる [阿部匡伸, 1995] [党建武・本多清志, 1995]。

本章の実験において、包絡線保存周波数と話者判定率の関係から、狭帯域包絡線に含まれる情報により話者知覚が可能であることがわかる。次章では、搬送波に含まれる情報を雑音と置き換え取り除いた信号を用い狭帯域包絡線のみの情報と話者知覚の関係を検討する。

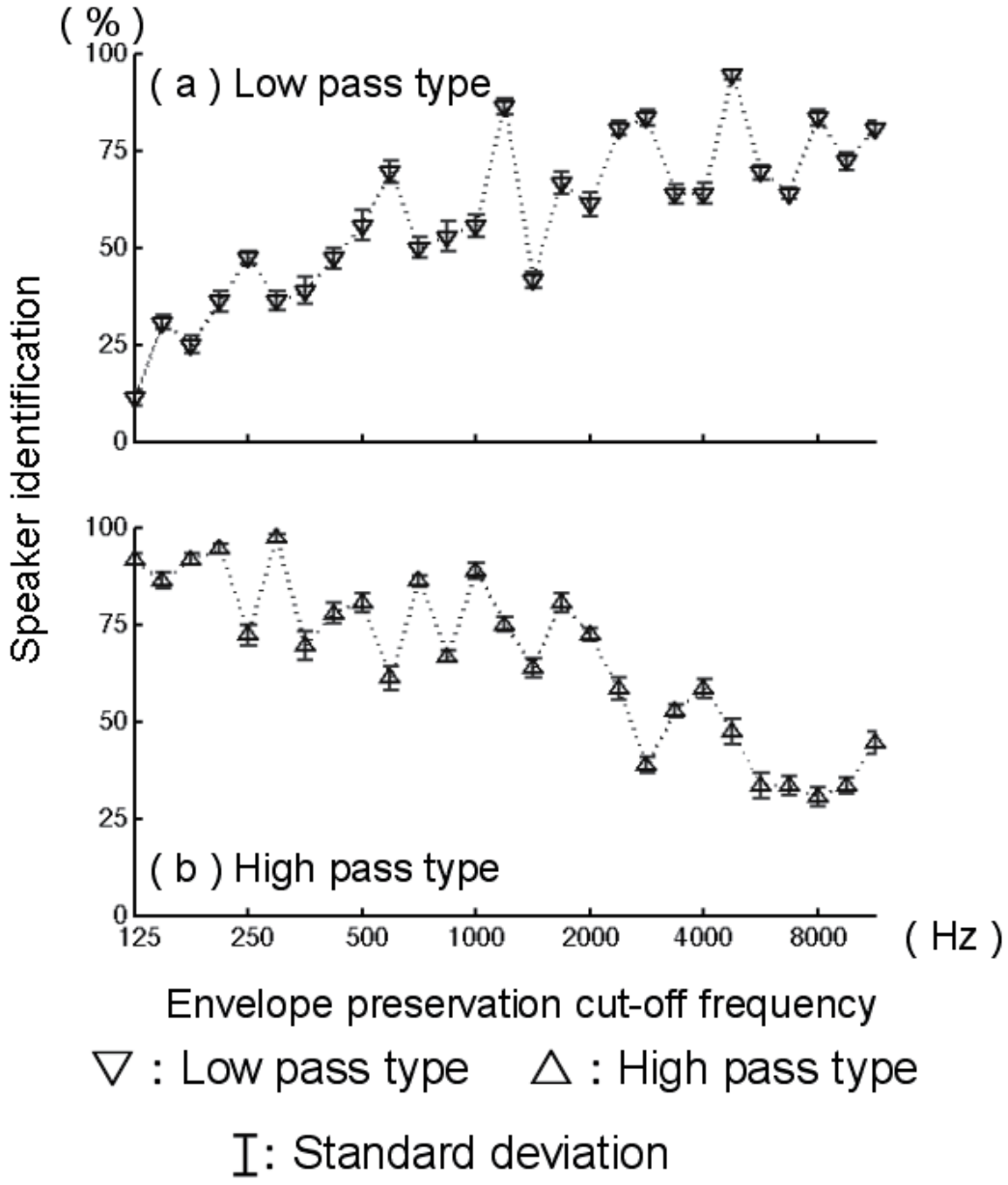

図 6.5: 女声合成信号試聴実験における話者判定率

6.3 帯域別包絡線変調雑音による話者判定実験

前節において狭帯域包絡線の保存帯域数と話者知覚の関係を述べた。これに対し、本節では帯域別包絡線変調雑音による試聴実験によりに狭帯域包絡線における話者情報を考察する。

6.3.1 試験音作成

Drullman は音声の了解性に関わる試聴実験において、1/4 オクターブ帯域波形包絡線と 1/4 オクターブ帯域雑音を合成する事により、搬送波の影響を取り除いた試聴実験を行った [Drullman, 1995] 。本試験は、この手法を用いて、2 章における合成信号の搬送波を帯域雑音の搬送波と置き換えることにより試験信号を作成した。即ち本試験信号は帯域別の雑音を音声の狭帯域包絡線により変調した包絡線変調雑音となる。試験音作成には男声 (25 名) を使用し、試験方法とサンプル数は 2 章と同様である。

6.3.2 狭帯域変調雑音を用いた試聴実験結果

図 6.4 と同様に図 6.6 に話者判定結果を示す。図 6.6a の LP 形の試験結果は搬送波が音声・話者情報を含まない雑音であっても狭帯域包絡線周波数の上昇に伴い話者判定率が上昇することを示した。また、図 6.6b の HP 形における試験結果は狭帯域包絡線周波数の上昇に伴い話者判定率が低下することを示した。この結果より、狭帯域包絡線の情報から話者知覚が可能であることがわかる。

図 6.7 に男声合成信号による話者判定率と狭帯域包絡線変調雑音による話者判定率の 2 次近似曲線を示す。図 6.7 における合成信号と包絡線変調雑音による話者判定率の推移は、話者の搬送波を用いた合成信号に対し、包絡線変調雑

音の判定率がやや低下しているものの、互いに同様な傾向を示している。 LP

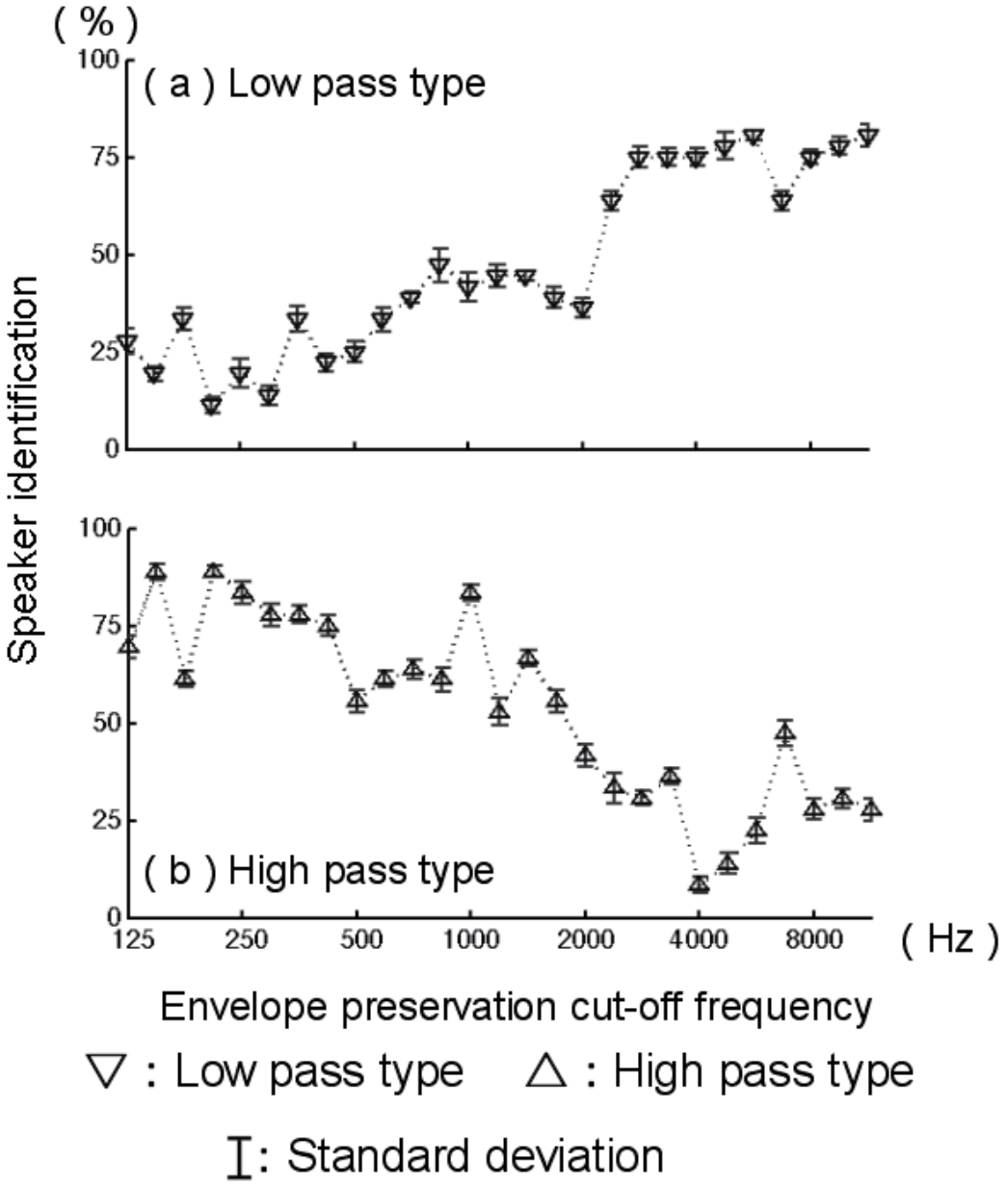

図 6.6: 狭帯域変調雑音試聴実験による話者判定率

形における話者判定率は話者情報に重要とされる第三ホルマントより低い周波数範囲 (2kHz 以下) の狭帯域包絡線のみでは低く、2kHz 以上の狭帯域包絡

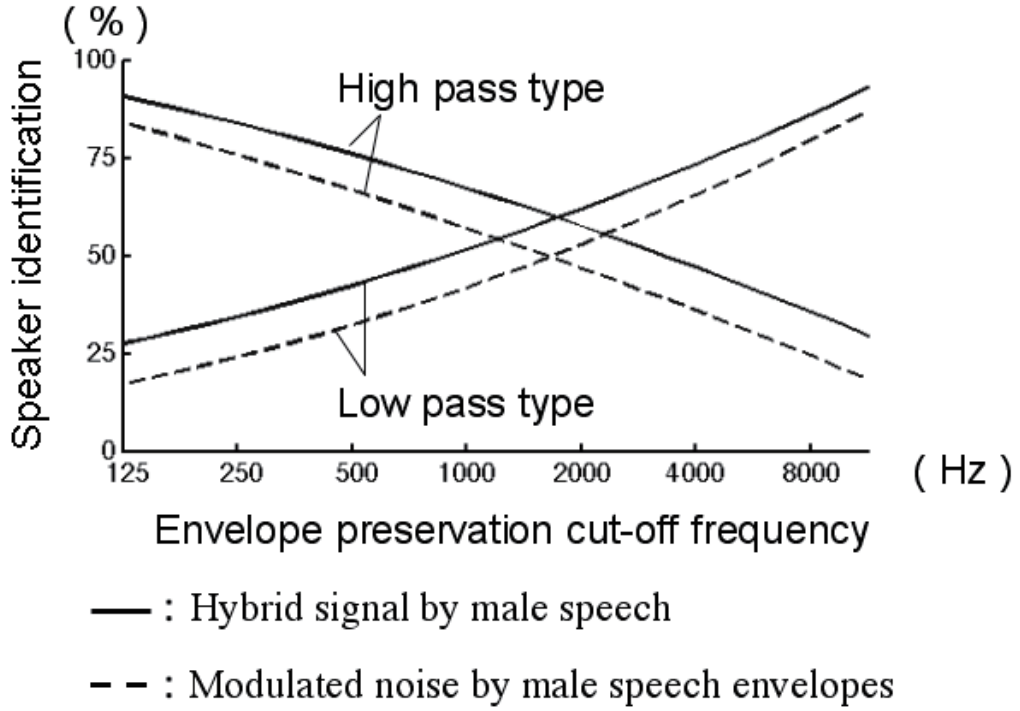

図 6.7: 男声合成信号 (図 6.4) と変調雑音 (図 6.6) による話者判定率の 2 次近似曲線

線情報を保存する事により話者判定が容易になる事がわかる。一方、HP 形における話者判定結果は第三ホルマント以上の狭帯域包絡線情報のみでは低く、F0 と関わる 500Hz 以下の狭帯域包絡線を保存すると話者判定が容易になる事がわかる。これは狭帯域信号の時間波形と関わる狭帯域包絡線が F0 推移の情報を保存していることが考えられる。これらの結果は、低域に含まれる F0 の推移情報の重要性と単母音における高域のスペクトル包絡における話者特徴両方が狭帯域包絡線による話者知覚においても重要であるを示している [Fitch, 1997] , [Van Dommelen, 1990] [北村達也・赤木正人, 1997] 。

6.4　振幅周波数特性と狭帯域包絡線

図 6.8 に原音声と包絡線変調雑音の各母音におけるスペクトル包絡を示す。また、各母音は試験信号に用いた一話者の音声信号「konojiha tooku kara mienikui」より切り出して用いた。また、帯域変調雑音スペクトルは帯域ごとに合成し算出したものである。図 6.8 において、変調雑音によるスペクトル包絡は原音声信号と近い傾向を有している事がわかる。さらに、個人差と関わるとされる下咽頭腔が影響する 2,500Hz 以上の周波数帯域も同様なスペクトルを有している事がわかる。この結果より、狭帯域包絡線はスペクトルの時間変化と母音に見られるような短時間スペクトルの情報を両方含み、音声の狭帯域包絡線による変調雑音においても話者知覚が可能なことを示した。

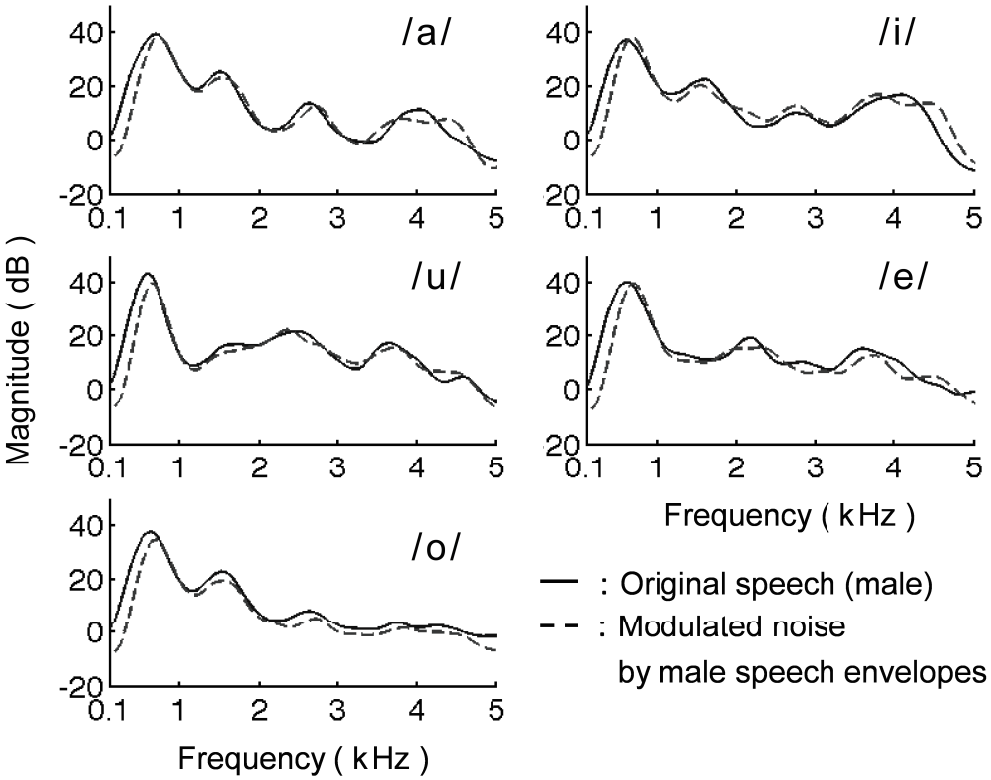

図 6.8: 変調雑音における母音部のスペクトル包絡分析

6.5　むすび

本章では音声信号において狭帯域包絡線が狭帯域搬送波にくらべ話者知覚に重要である事を明らかにした。さらに、帯域別搬送波を帯域雑音に置き換えた狭帯域包絡線変調雑音によっても話者判定が可能な事から、音声の帯域別包絡線に話者情報が含まれることを示した。狭帯域包絡線による話者知覚において話者判定率は狭帯域包絡線保存数に比例して上昇し、低域と高域両方の包絡線情報を保存することにより話者知覚が容易になることを示した。また、狭帯域包絡線における話者情報は、話者知覚に重要な F0 の推移情報等や第三ホルマントなどの高域の周波数情報を含み、低域から高域に広くわたる狭帯域包絡線により表現可能であることを示した。これらの結果より、狭帯域包絡線により音声の言語情報だけでなく話者情報も知覚可能であることを示した。

第7章　狭帯域包絡線相関行列を用いた話者特徴表現

7.1　まえがき

前章において狭帯域包絡線が話者情報を含んでいることを示した。本章では前章の応用として音声における狭帯域包絡線の帯域間相関による話者の個人特徴表現について検討する。

話者照合や識別は音声分析技術において魅力的な応用分野である [Quatieri et al., 2000] [Zilovic et al., 1998] [Bachorowski and Owren, 1999]。Bimbot 等 [Bimbot et al., 1995] は 4 - 8 kHz の電話帯域より高い短時間パワースペクトルの時間変化に話者特徴が含まれることを示唆した。そこで狭帯域包絡線の帯域間相関が短時間パワースペクトルの時間変化と関係する [Par and Kohlrausch, 1998] ことから、音声の個人特徴を表現できると考えた。

本研究では男声 25 名、女声 28 名による包絡線相関行列（ECM:envelope correlation matrices）の類似度を用いた話者識別実験を行った。ECM は全て同一文章の発話による音声信号により作成した。また音声信号は電話帯域音声（250 Hz − 3 kHz)、高周波数帯域音声（2-11.3 kHz)、広周波数帯域音声（250 Hz -11.3 kHz）の異なる周波数範囲の信号を用いて実験を行った。その結果、ECM のを用いて話者識別可能であることを示す。

7.2　狭帯域包絡線分析

話者特徴が短時間パワースペクトラムの時間変化により表現できることから、本節では狭帯域包絡線の相関行列による話者特徴の表現を考察する。狭帯域包絡線の相関行列作成にあたり音声信号を収音した。被験者は全て母国語が日本語、それぞれの話者には同一の文章を発話させサウンドブースで収音 (サンプリング周波数４８ kHz １６ bit の A/D 変換機を用いた。) を行った。全ての被験者は同一文章について３回づつの発話を収音した。音声信号はそれぞれ 2 - 11.3 kHz の 21 個の 1/8 オクターブバンドと 250 Hz-2 kHz の１２個の 1/4 オクターブバンド帯域に計 33 帯域に分割した。それぞれの帯域包絡線は半波整流後 40Hz のローパスフィルターに通し得た。この帯域包絡線を用い、それぞれの話者の包絡線帯域間相関行列 (33×33) を作成した。包絡線相関行列は以下で定義される

$$\rho(i,j) = \frac{\overline{E_i(n)E_j(n)}}{\sqrt{\overline{E_i^2(n)}\ \ \overline{E_j^2(n)}}} \tag{7.1}$$

ここで

$$E_l(n) = E_{l_0}(n) - \overline{E_{l_0}(n)}. \tag{7.2}$$

$\bar{*}$ は長時間平均、$E_{l_0}(n)$ は l 番目の帯域包絡線、n はサンプルを表す。本研究では２種類の包絡線相関行列をそれぞれの話者毎に作成した。一つ（reference-ECM）は話者特徴のテンプレートとして 2 発話を用いて作成した包絡線相関行列。もう一つ（test-ECM）は話者照合用に環境雑音を含む一つの発話を用いて作成した包絡線相関行列である。

7.3 ECMによる話者識別実験

ECM に含まれる話者特徴を明らかにするため図 7.1 に示すような ECM を用いた雑音下における話者識別実験を行った。図 7.2 に環境雑音の振幅スペク

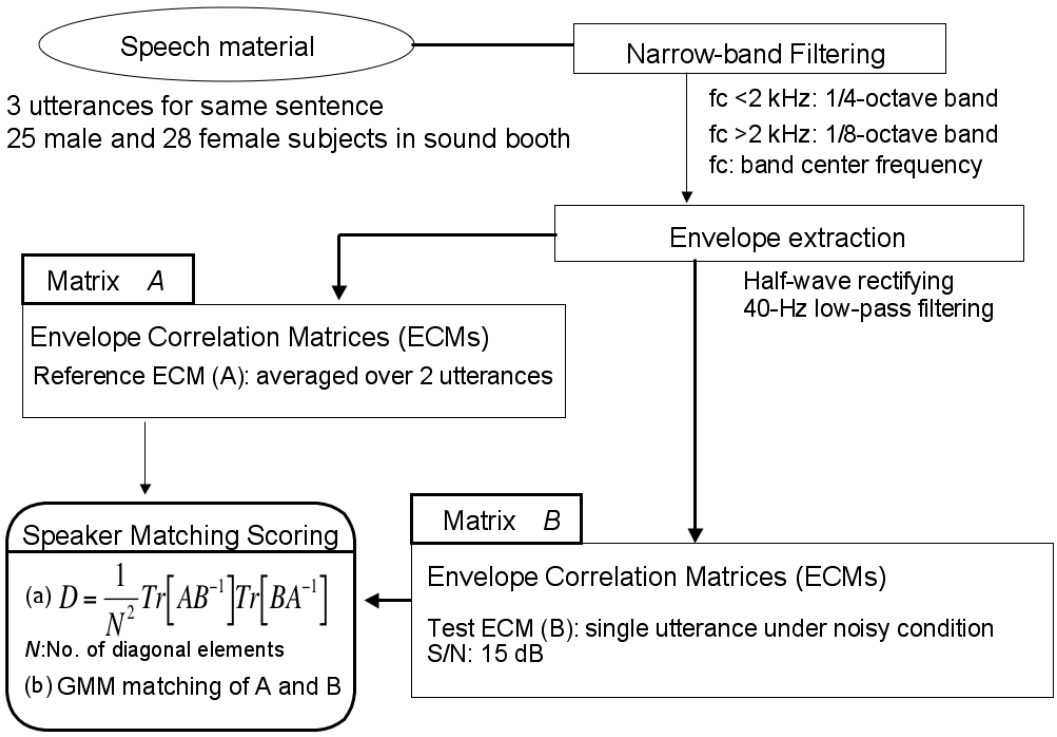

図 7.1: 話者識別実験方法

トルを示す。図 7.3 は話者の reference-ECM のサンプルである。

図 7.3 より reference-ECM に話者の違いがあらわれていることがわかる。test-ECM と reference-ECM の類似度は harmonic sphericity measure により求められる [Bimbot et al., 1995] [Zilca, 2002]。

$$D = \frac{1}{N^2} \mathrm{T_r}[AB^{-1}] \mathrm{T_r}[BA^{-1}] \tag{7.3}$$

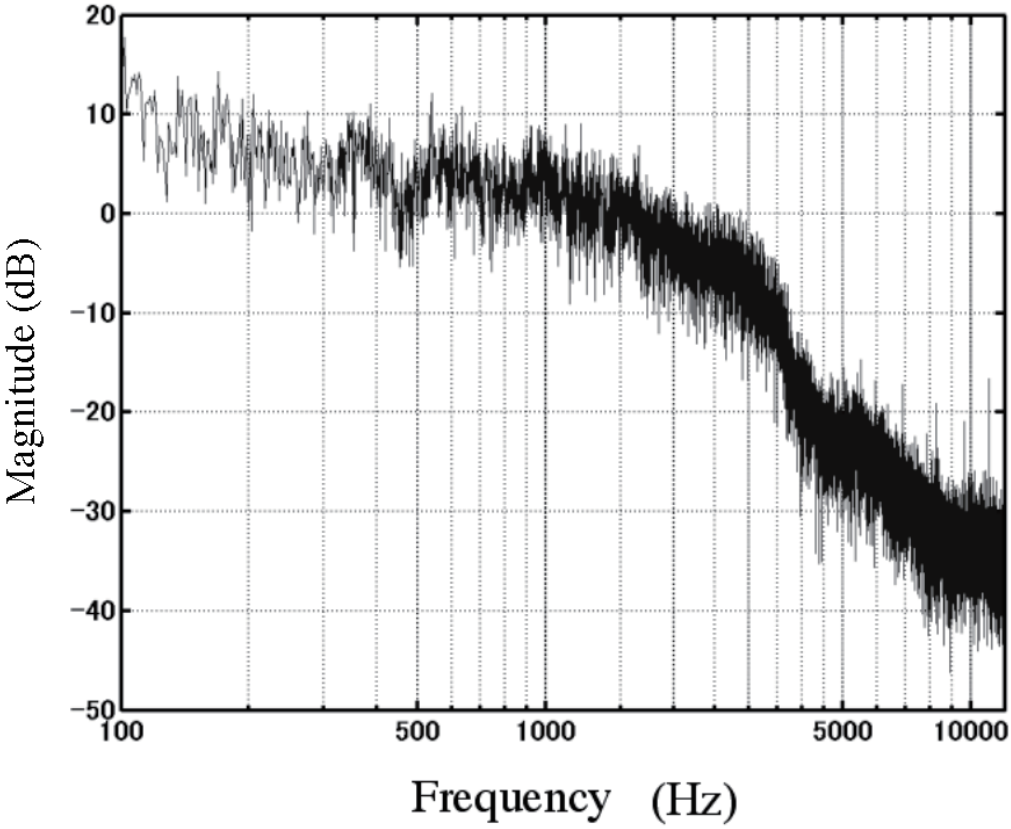

図 7.2: 環境雑音の振幅スペクトル

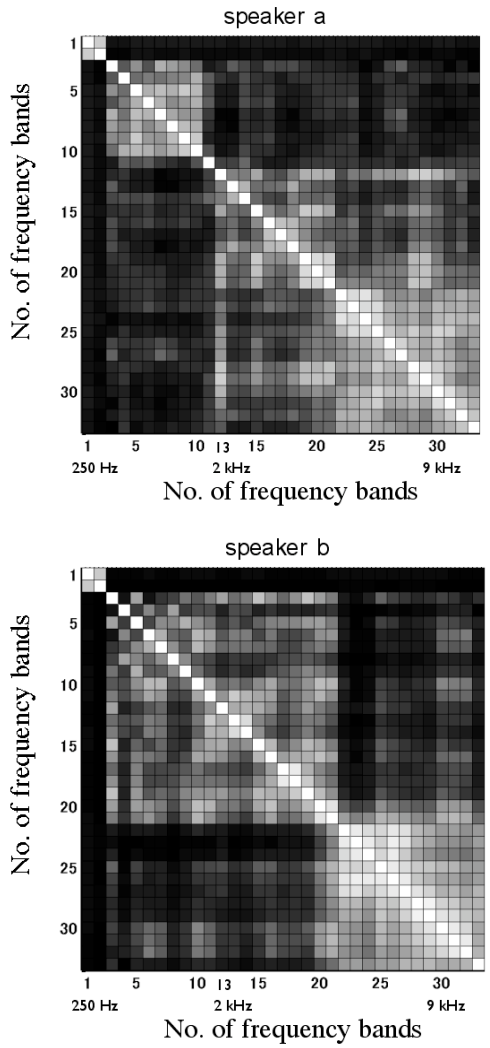

図 7.3: reference ECM の例

ここで A と B は reference-ECM と test-ECM をそれぞれ表す。また、N は相関行列の対角要素数である。また、本実験では Gaussian Mixture Model を用いた話者識別も行った [Campbell, 1997]。GMM による話者識別実験いおいて reference-ECM 用の 2 つの発話のうち一つを学習用に用いた。識別においてマッチングスコアには cohort normalization を用いた [Rosenberg et al., 1992]。

マッチングテストでは対数尤度によるマッチングスコアが出力される。正規化されていな識別スコアは

$$score_u = \log \widehat{p}(\mathbf{O}|I) \tag{7.4}$$

となり、$\widehat{p}$ は尤度を $\mathbf{O}$ は ECM の列ベクトルの配列を I は話者番号を表す。このとき cohort nomalized score は対数尤度の差により定義できる。

$$score_n = \log \widehat{p}(\mathbf{O}|I) - \max_k[\log \widehat{p}(\mathbf{O}|c_k(I))] \tag{7.5}$$

ここで、$\log \widehat{p}(\mathbf{O}|c_k(I))$ は $C(I)$ と話者 I と k 番目のモデルの観測ベクトル列の対数尤度を示す。図 7.4 は D^{-1}（左のパネル）と GMM（右のパネル）によるマッチングスコアを示す。マッチングテストは男声と女声で別々に行った。D^{-1} では女声の識別において 1 名識別ミスがあるが、ECM が話者の個人情報を含み話者識別可能であることがわかる。

7.4　異なる ECM の周波数範囲における話者識別

短時間パワースペクトルの時間変化に基づく話者特徴分析は高い周波数を含む広い周波数範囲が必要となる。そこで 3 種類の低周波数帯域音声（250 - 2 kHz)、高周波数帯域音声（2 - 11.3 kHz)、電話帯域音声（250 Hz - 3 kHz)、の異なる周波数範囲にわたる ECM を用意し帯域別話者特徴分析を行った。図 7.5 はそれぞれの ECM による識別結果を示す。パネル (a) は GMM と cohort normalization による低周波数範囲の結果を、パネル (b) は高周波数範囲の結果を、(c) は電話帯域による試験結果を示す。その結果、高い周波数範囲を含む ECM が音声の識別に有用であることがわかった。

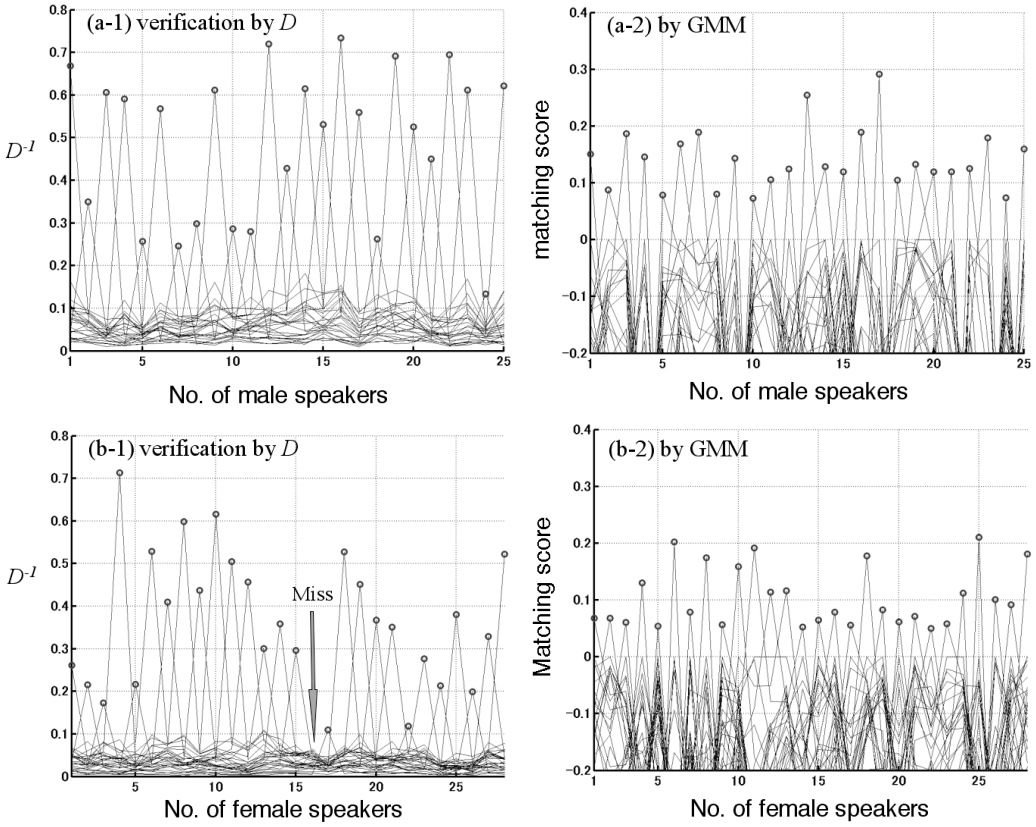

図 7.4: 広周波数帯域 (250 Hz - 11.3 kHz) を用いた話者識別実験結果 (a) 男声 (b) 女声

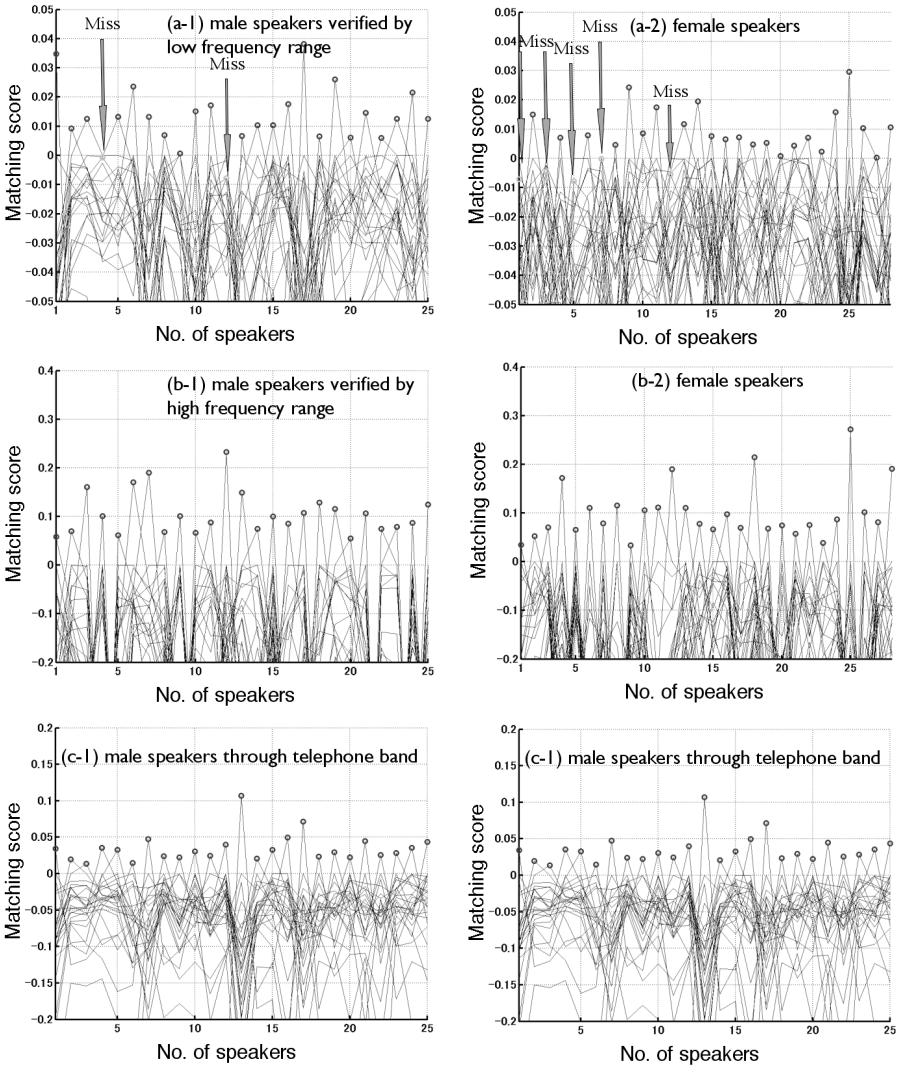

図 7.5: 周波数帯域と話者識別結果; (a) 低周波数範囲 (fc $<$ 2 kHz), (b) 高周波数範囲 (fc $>$ 2 kHz), (c) 電話帯域 (250 Hz$<$ fc $<$ 3 kHz)

7.5 ECM の間引きを用いた話者識別

これまで ECM を用いた話者特徴表現を考察した。本節では一つの ECM の要素を間引くことにより異なる ECM を作成し識別率の向上を試みる。図 7.6 に間引きを用いた話者識別方法を示す。ECM の間引きによる話者識別は隣り合う相関行列要素を取り出し異なる 2 つの ECM を作成し、それぞれの ECM より出された ECM 距離を平均することにより話者識別を行う。話者識別は ECM の

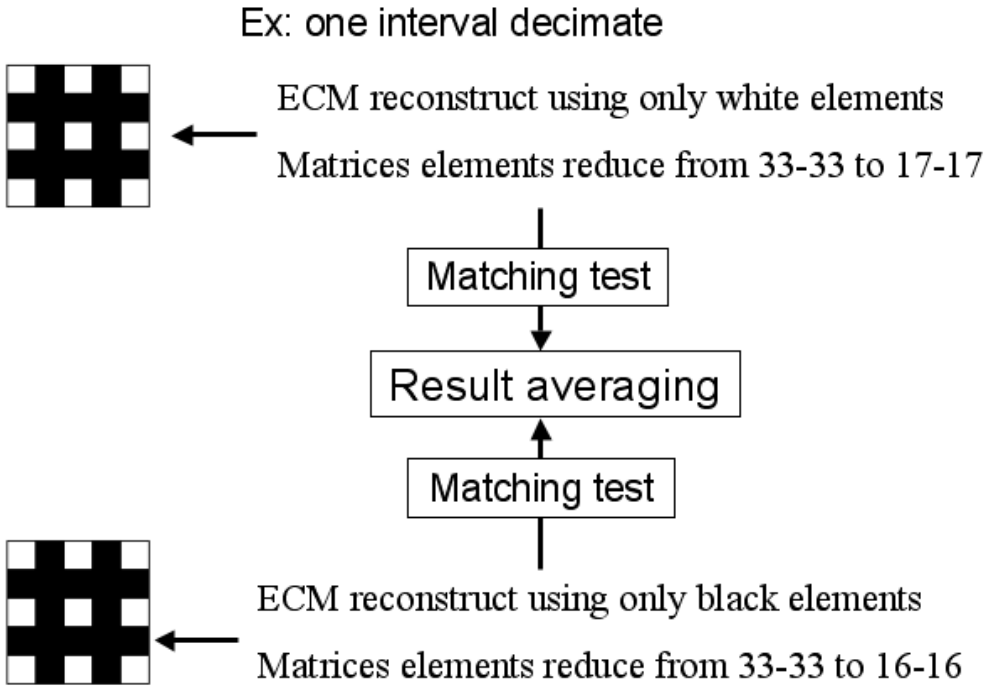

図 7.6: ECM の間引きを用いた話者識別

harmonic sphericity measure による類似度を用いて評価した。また、本実験は図 7.4 と同様の信号を用いて評価することとした。広周波数帯域における話者識別結果を図 7.7 に示す。図において 図は (a) 男声の ECM、(b) 女声の ECM、(c) 男声の間引きを用いた ECM、(d) 女声の間引きを用いた ECM の識別結果を示す。この結果、間引きを用いた話者識別では識別結果が向上しているこ

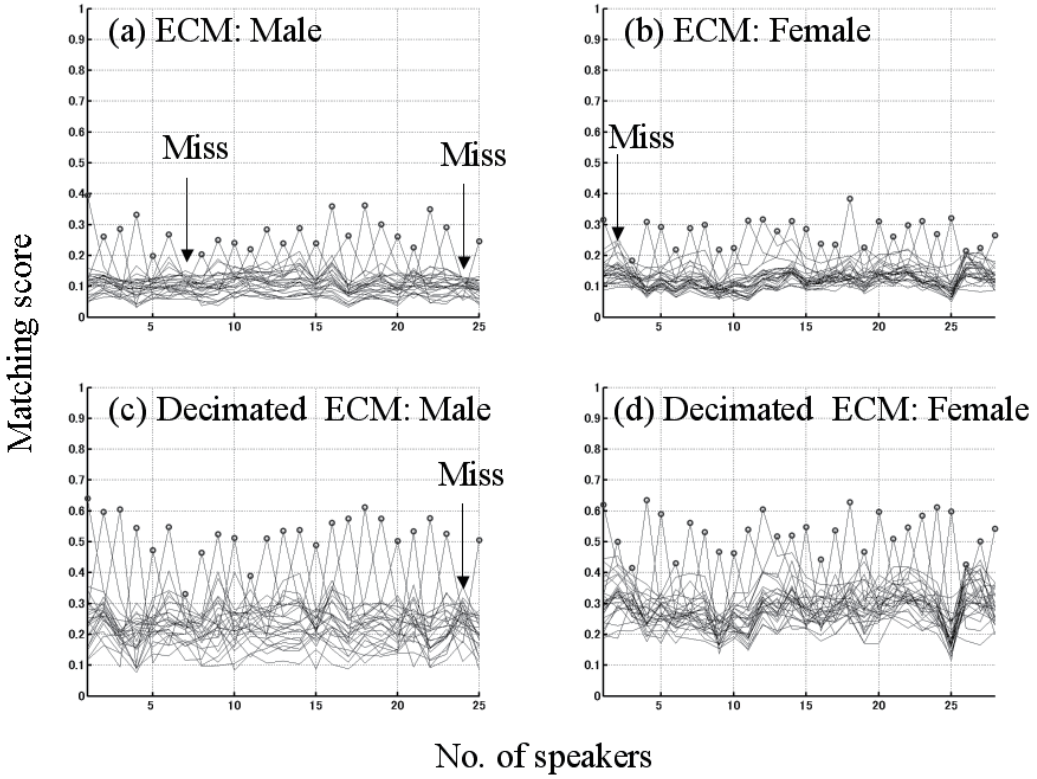

図 7.7: 間引きを用いた ECM と話者識別結果

とが解る。本節では ECM の間引きを用いることにより識別率の向上が図れることを明らかにした。次節では話者識別における従来法である Mel-Frequency Cepstrum Coffiecient を用いた Gaussian Mixture Model による話者識別結果と ECM による比較を行う。

7.6 ECM の話者識別率

本節では従来より用いられる Mel-Frequency Cepstrum Coffiecient(MFCC) を用いた Gaussian Mixture Model(GMM) における話者識別率と ECM の識別率の比較を行う。本実験では前節で使用した男声の音声信号を用いて話者識別を行った。さらに、実験信号における音声信号は 2s、1s、0.3s の三つを用い、信号の時間長と識別率変化を考察した。

ECM の話者識別は同様の文章を発話した 3 文章を用い、2 発話を reference-ECM 用、1 発話を test-ECM 用に用い話者識別実験を行った。MFCC を用いた GMM による話者識別は、各信号を 20ms にフレーム分割し MFCC を求め、1 発話を reference 製作にもう 1 発話を Em-traning に用い、残りの 1 発話を test 用とした。

本実験の結果、ECM の話者識別率が信号長 3 s において 100 %、1s において 100 %、0.3s において 80 %となった。また、従来法である MFCC による識別率は 3 s において 96 %、1s において 96 %、0.3s において 80 %となった。この結果、ECM による話者識別法は同一文章を用いた識別にキーワードを用いない話者識別において従来法である MFCC の GMM と比較し良い識別率となった。

7.7 考察

本章では前章より明らかとなった狭帯域包絡線と話者情報の関係に着目し、狭帯域包絡線の帯域間相関行列による話者特徴表現を試みた。その結果、帯域

相関行列を用いて話者識別が可能であることを示した。さらに、言語情報によらない同一文章を用いた話者識別実験において従来法である MFCC による GMM を用いた話者識別に対して帯域間相関行列を用いた話者識別では識別率が向上することを示している。このことより、狭帯域包絡線における話者情報は狭帯域包絡線の帯域間情報の関係により知覚していることが考えられる。

7.8　むすび

本研究では時間波形の狭帯域包絡線の帯域間相関による話者特徴表現を試みた。２種類の包絡線相関行列 (ECM:envelope correlation matrices) を５３名の話者の音声信号より作成し、２つの発話から作成する話者情報の reference-ECM に対し、１つの発話より作成した test-ECM が reference-ECM と同一話者かを判定した。その結果、ECM により話者の個人特徴が識別可能であり電話帯域音声であっても識別できることを示した。また、ECM の間引きを用いた話者識別により話者識別精度を向上することがわかった。さらに、ECM を用いた話者識別法と従来より話者識別に用いられる MFCC による GMM の話者識別法を比較した結果、MFCC に比べ ECM の話者識別率が高いことが解った。これより、前章で示した狭帯域包絡線に含まれる話者情報から話者識別が可能であることを示した。

第8章　総括

結論

本研究は、従来音源の特徴表現にあまり用いられていない位相スペクトルに着目し、音源情報知覚と位相スペクトルの関係を検討したものである。第２章において、位相情報と音声了解性の関係を音声了解度評価試験を行い考察した。試聴実験において 1/16ms - 2048ms のフレーム長を用いた短時間フーリエ変換による音声信号の振幅情報と位相情報を白色雑音による情報と入れ替えた合成信号を用いることにより音声了解度を調べた。その結果、256ms より長いフレーム長と 4ms より短いフレーム長において位相情報が音声了解度に重要であることを明らかにした。さらに位相情報が狭帯域包絡線の復元に重要であることを示した。また、この実験結果よりフーリエ変換において知られる時間分解能と周波数分解能のトレードオフの関係に対し、音声情報では振幅情報と位相情報による音声了解性のトレードオフが生じる事を示した。これにより、音源情報と関わる位相情報の変化を狭帯域包絡線を通じて知覚する可能性を示唆した。　第３章では、音声了解度に位相情報が重要となることから、音声了解度が低い場合の評価方法である音声明瞭度評価と位相情報の関係について考察した。試聴実験において音声と無相関の雑音をもちいた雑音付加信号の音声了解度を単音節明瞭度試聴実験により評価した。さらに、音声明瞭度評価における従来法との比較において MTF-STI を用いた。その結果、MTF-STI と PCI による評価が高い相関関係にあり音声明瞭度が位相スペクトルを用いて評

価可能であることを示した。これより音声情報知覚に狭帯域包絡線と関わる位相情報が大きく関係することを明らかにした。

第 4 章では、音声了解度が高い場合によく用いらる音声品質評価と狭帯域包絡線の関係について試聴実験を行い考察した。その結果、PESQ 値と比較し Opinion 評価結果が低い場合においても予測が可能であることが解った。これより、狭帯域包絡線の振幅ヒストグラムと音声信号品質評価に相関があることから、狭帯域包絡線に音声明瞭度だけでなく品質評価に関わる情報も含まれることを示した。

第 5 章では、音源と相関のある信号が含まれる場合の音声了解度試聴実験を行った。さらに、音源情報と狭帯域包絡線の関係を信号の類似性に基づき考察した。狭帯域包絡線に音源情報が含まれる場合、信号情報に基づく類似性は時間波形における振幅変化で評価できると考えた。そこで試聴実験において情報マスキング効果に着目した信号類似度評価をおこなった。その結果、信号の振幅分布から情報マスキングに関わる信号類似度が評価可能であることを示し、狭帯域包絡線に音声以外の音源特徴も含まれることを明らかにした。

第 6 章では、前章において明らかとなった狭帯域包絡線における音源特徴に着目し、狭帯域包絡線と音声に含まれる話者情報の関係について考察した。Li 等 (1974) はスペクトルの時間変化による話者特徴分析により狭帯域包絡線に含まれる話者情報に言及したが、狭帯域包絡線と話者情報の知覚の関係は示さなかった。そこで音声信号における狭帯域包絡線と狭帯域搬送波を入れ替えた合成信号による話者判定試聴実験を行った。その結果、狭帯域包絡線を話者情報知覚の手掛かりにしていることが明らかになった。

第 7 章では、第 6 章で示した狭帯域包絡線における話者情報を帯域間相関行列を用いて表現し話者識別実験を行った。その結果、従来法における MFCC による GMM を用いた話者識別結果と比較し、話者識別率が従来法に比べ向上

することを示した。

本論文は音源信号の特徴をその位相特性の変化に見いだし、位相特性が音源情報知覚さらには音源識別に及ぼす影響効果について研究したものである。その結果、狭帯域包絡線に含まれる音源情報の所在として音声了解性に関わる情報は狭帯域包絡線類似度、音質に関わる情報は狭帯域包絡線の振幅ヒストグラム、話者特徴は狭帯域包絡線の帯域間類似度に現れることを示した。以上より、従来音響信号の特徴表現に一般的に用いない位相情報に音源情報が包括的に含まれ、その位相情報における音源情報は狭帯域包絡線を通じて知覚することを明らかにした。

今後の課題

本論文では位相情報に含まれる音源情報が狭帯域包絡線を通じ知覚されることを示した。狭帯域包絡線が聴知覚において利用される事は近年明らかになりつつある。今後は狭帯域包絡線の分析と知覚の関係を示すことが必要となる。狭帯域包絡線の情報は、現在、人工蝸牛等の技術に用いられているが、高精度な人工蝸牛の開発には音源特徴と狭帯域包絡線の関連を信号処理に着目し研究する必要がある。このような研究を行うことにより、音源情報知覚と音響信号の物理的特徴の従来の周波数分析と異なる新たな信号分析法が確立される可能性がある。本論文において示した音源情報は音声というコミュニケーションにおいて非常に重要な情報ではあるが、音声以外の音源に関わる知覚の関係は未だ課題として残る。本論文では、位相情報が波形を構成する事から、聴覚において利用される狭帯域包絡線が位相情報から構成され知覚されることを明らかにした。今後は、聴知覚において音源情報を抽出する信号処理手法、さらには多様な音源情報と知覚の関係を明らかにすることが期待される。

参考文献

Arbogast, T.L., C.R. Mason, and G. Kidd Jr (2002) "The effect of spatial separation on informational and energetic masking of speech," *The Journal of the Acoustical Society of America*, Vol. 112, pp. 2086–2098.

Bachorowski, J.A. and M.J. Owren (1999) "Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech," *The Journal of the Acoustical Society of America*, Vol. 106, pp. 1054–1063.

Beerends, J.G., A.P. Hekstra, A.W. Rix, and M.P. Hollier (2002) "Perceptual Evaluation of Speech Quality (PESQ): The New ITU Standard for End-to-End Speech Quality Assessment Part II-Psychoacoustic Model," *JOURNAL-AUDIO ENGINEERING SOCIETY*, Vol. 50, No. 10, pp. 765–778.

Bimbot, F., I. Magrin-Chagnolleau, and L. Mathan (1995) "Second-order statistical measures for text-independent speaker identification," *Speech Communication*, Vol. 17, No. 1-2, pp. 177–192.

Boll, S. (1979) "Suppression of acoustic noise in speech using spectral subtraction," *Acoustics, Speech, and Signal Processing [see also IEEE Transactions on Signal Processing], IEEE Transactions on*, Vol. 27, No. 2, pp. 113–120.

Campbell, J.P. (1997) “Speaker Recognition: A Tutorial,” *Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1437–1462.

Drullman, R. (1995) “Temporal envelope and fine structure cues for speech intelligibility,” *The Journal of the Acoustical Society of America*, Vol. 97, pp. 585–592.

Drullman, R. and A.W. Bronkhorst (2004) “Speech perception and talker segregation: Effects of level, pitch, and tactile support with multiple simultaneous talkers,” *The Journal of the Acoustical Society of America*, Vol. 116, pp. 3090–3098.

Durlach, N.I., C.R. Mason, B.G. Shinn-Cunningham, T.L. Arbogast, H.S. Colburn, and G. Kidd Jr (2003) “Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity,” *The Journal of the Acoustical Society of America*, Vol. 114, pp. 368–379.

Espinoza-Varas, B. and SV Cherukuri (1995) “Evaluating a model of auditory masking for applications in audiocoding,” *Applications of Signal Processing to Audio and Acoustics, 1995., IEEE ASSP Workshop on*, pp. 195–197.

Fitch, W.T. (1997) “Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques,” *The Journal of the Acoustical Society of America*, Vol. 102, pp. 1213–1222.

Freyman, R.L., K.S. Helfer, D.D. McCall, and R.K. Clifton (1999) “The role of perceived spatial separation in the unmasking of speech,” *The Journal of the Acoustical Society of America*, Vol. 106, pp. 3578–3588.

Gotoh, S., M. Kazama, M. Tohyama, and Y. Yamasaki (2006) "Speaker Verification Using Narrow-band Envelope Correlation Matrices," in *2006 IEEE International Symposium on Signal Processing and Information Technology*, pp. 310–313.

Hoen, M., F. Meunier, C.L. Grataloup, F. Pellegrino, N. Grimault, F. Perrin, X. Perrot, and L. Collet (2007) "Phonetic and lexical interferences in informational masking during speech-in-speech comprehension," *Speech Communication*, Vol. 49, No. 12, pp. 905–916.

Houtgast, T. and H.J.M. Steeneken (1973) "The Modulation Transfer Function in Room Acoustics as a Predictor of Speech Intelligibility," *The Journal of the Acoustical Society of America*, Vol. 54, p. 557.

——— (1985) "A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria," *The Journal of the Acoustical Society of America*, Vol. 77, pp. 1069–1077.

Houtgast, T., H.J.M. Steeneken, and R. Plomp (1980) "Predicting speech intelligibility in rooms from the modulation transfer function. i. general room acoustics," *Acustica*, Vol. 46, No. 60-72.

Li, K.P. and GW Hughes (1974) "Talker differences as they appear in correlation matrices of continuous speech spectra," *The Journal of the Acoustical Society of America*, Vol. 55, pp. 833–837.

Liu, L., J. He, and G. Palm (1997) "Effects of phase on the perception of intervocalic stop consonants," *Speech Communication*, Vol. 22, No. 4, pp. 403–417.

Nakashima, N., T. Tagaeto, M. Tohyama, and RH Lyon (1996) “Cepstrum deconvolution for estimating probability density functions of compound signals,” *Proceedings of Internoise*, pp. 2821–2824.

Navarro, M.P.N. and R.L. Pimentel (2007) “Speech interference in food courts of shopping centres,” *Applied Acoustics*, Vol. 68, No. 3, pp. 364–375.

Oppenheim, AV and JS Lim (1981) “The importance of phase in signals,” *Proceedings of the IEEE*, Vol. 69, No. 5, pp. 529–541.

Van de Par, S. and A. Kohlrausch (1998) “Analytical expressions for the envelope correlation of narrow-band stimuli used in CMR and BMLD research,” *The Journal of the Acoustical Society of America*, Vol. 103, pp. 3605–3620.

Quatieri, TF, DA Reynolds, and GC O’leary (2000) “Estimation of handset nonlinearity with application to speakerrecognition,” *IEEE Transactions on Speech and Audio Processing*, Vol. 8, No. 5, pp. 567–584.

Rhebergen, K.S., N.J. Versfeld, and W.A. Dreschler (2005) “Release from informational masking by time reversal of native and non-native interfering speech,” *The Journal of the Acoustical Society of America*, Vol. 118, pp. 1274–1277.

Rix, AW, JG Beerends, MP Hollier, AP Hekstra, and I. PsyTechnics (2001) “Perceptual evaluation of speech quality (PESQ)-a new method forspeech quality assessment of telephone networks and codecs,” in *Acoustics, Speech, and Signal Processing, 2001. Proceedings.(ICASSP’01). 2001 IEEE International Conference on*, Vol. 2.

Rosenberg, A.E., J. DeLong, C.H. Lee, B.H. Juang, and F.K. Soong (1992) "The Use of Cohort Normalized Scores for Speaker Verification," in *Second International Conference on Spoken Language Processing*, ISCA.

Schmitz, CD and N. Iyer (2003) "On the reduction of masking effects while preserving competing binaural audio streams," *Signals, Systems and Computers, 2003. Conference Record of the Thirty-Seventh Asilomar Conference on*, Vol. 1, pp. 740–744.

Schroeder, M.R. (1999) *Computer speech*: Springer-Verlag Berlin Heidelberg, pp.63–73.

Schroeder, M.R. and H.W. Strube (1986) "Flat-spectrum speech," *J. Acoust. Soc. Am*, Vol. 79, No. 5, pp. 1580–1583.

Shannon, R.V., F.G. Zeng, V. Kamath, J. Wygonski, and M. Ekelid (1995) "Speech Recognition with Primarily Temporal Cues," *Science*, Vol. 270, No. 5234, pp. 303–304.

Smith, Z.M., B. Delgutte, and A.J. Oxenham (2002) "Chimaeric sounds reveal dichotomies in auditory perception," *Nature*, Vol. 416, pp. 87–90.

Traumueller, A. and M.E.H. Schouten (1987) *The Psychophysics of Speech Perception*: Kluwer Academic Print on Demand, pp.1332–1350.

Van Dommelen, W.A. (1990) "Acoustic parameters in human speaker recognition," *Language and Speech*, Vol. 33, No. 3, pp. 259–272.

Van Engen, K.J. and A.R. Bradlow (2007) "Sentence recognition in native-

and foreign-language multi-talker background noise," *The Journal of the Acoustical Society of America*, Vol. 121, pp. 519–526.

Vary, P. (1985) "Noise suppression by spectral magnitude estimation-mechanism and theoretical limits," *Signal processing*, Vol. 8, No. 4, pp. 387–400.

Wang, C. and JS Bradley (2002) "Prediction of the speech intelligibility index behind a single screen in an open-plan office," *Applied Acoustics*, Vol. 63, No. 8, pp. 867–883.

Wilson, B.S., C.C. Finley, D.T. Lawson, R.D. Wolford, D.K. Eddington, and W.M. Rabinowitz (1991) "Better speech recognition with cochlear implants," *Nature*, Vol. 352, pp. 236–238.

Zilca, RD (2002) "Text-independent speaker verification using utterance level scoring and covariance modeling," *Speech and Audio Processing, IEEE Transactions on*, Vol. 10, No. 6, pp. 363–370.

Zilovic, MS, RP Ramachandran, RJ Mammone, and R.B. Bellcore (1998) "Speaker identification based on the use of robust cepstral featuresobtained from pole-zero transfer functions," *Speech and Audio Processing, IEEE Transactions on*, Vol. 6, No. 3, pp. 260–267.

阿部匡伸 (1995) 「基本スペクトルと漸次変形による音声モーフィング」,『音講論集』, 259–260 頁.

党建武・本多清志 (1995) 「基母音発生時の音声スペクトルに対する梨状窩の影響」,『信学技報』, SP95-10 頁.

風間道子・東山三樹夫・山崎芳男 (2009) 「狭帯域音声波形包絡線の帯域間相関行列に現れる話者情報」,『電子情報通信学会論文誌 A』, 第 92 巻, 第 4 号, 205–215 頁.

北村達也・赤木正人 (1997) 「単母音の話者識別に寄与するスペクトル包絡成分」,『日本音響学会誌』, 第 53 巻, 第 3 号, 185–191 頁.

研究業績

論文／査読付国際会議

- "Using HDC to evaluate signal similarity for information masking," Applied Acoustics, Vol.70(5) pp689-694, Satoru Gotoh, Mitsuo Matsumoto and Yoshio Yamasaki

- "On the Significance of Phase in the Short-Term Fourier Spectrum for Speech Intelligibility," Journal of Acoustical Society of America, (2009年12月採録決定), Michiko Kazama, Satoru Gotoh, Mikio Tohyama, and Tammo Houtgast

- "Speaker Verification Using Narrow-band Envelope Correlation Matrices," IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) 2006 Vancouver, Canada, Satoru Gotoh, Michiko Kazama, Mikio Tohyama and Yoshio Yamasaki

- "Speech Intelligibility Estimation Based on Phase Correlation Index," Measurement of Speech and Audio Quality in Network (MESAQUIN) 2005, Prague Czech Republic, Satoru Gotoh, Michiko Kazama and Mikio Tohyama

国際会議

- ”Reconstruction of missing formants based on spectral power series expansion,” 156th meeting: Acoustical Society of America,2008, Vol.124,p.2578, Kesaaki Minemura, Satoru Goto and Mikio Tohyama

国内学会

- ”位相情報による音声信号表現,”日本音響学会秋季研究発表会,2004 年, 3-Q-26, 後藤理, 風間道子, 東山三樹夫, 篠原克幸

謝辞

本研究の遂行にあたり、私を研究室の博士課程学生として受け入れて頂き、終始適切な御指導と御議論を頂いた早稲田大学の山崎芳男教授と音響情報処理研究室の皆様に感謝致します。本研究の端緒と着想をいただき、私が工学院大学に所属していた頃から、研究方針・発表技術・論文作成などあらゆる面で御指導を頂いた早稲田大学の東山三樹夫教授に感謝致します。さらに、本論文作成に当たり、貴重な御意見、御指導いただきました、大谷淳教授、河合隆史教授、及川靖広准教授に感謝いたします。

本論文の信号処理における計算手法について適切なご指導をいただいたオランダの VU university の Tommo Houtgast 教授、工学院大学の高橋静昭教授ならびに三好和憲教授に感謝致します。論文の執筆及び実験に多大な協力を頂いた、松本光雄氏、風間道子氏、早稲田大学白井克彦研究室谷口徹氏、宮島崇浩氏、久保陽太郎氏、峰村今朝明氏に感謝致します。研究生活を支えて頂いた音響情報処理研究室の先輩方である、大内康裕氏、小西雅氏、中沢誠氏、ビッグラブの池田雄介氏、音響研究室表現工学科一期生に感謝致します。さらに、音研水泳部では研究を超えた精神・肉体面をご指導して頂きました。武岡成人氏には鋼鉄のような精神力、野口紗生氏には水と人とのあり方を再考するきっかけを作って頂きました。さらに、共に泳いだ神保直史氏、酒井寿理氏、石井紀義氏、中嶌仁彦氏、原川泰紀氏、研究面においても多大にご協力頂いた八十島乙暢氏に充実した研究生活を支えて頂き感謝致します。

本研究を進めるにあたり、日頃から叱咤激励を頂いたリオン株式会社藤坂洋

一氏、奥野貴俊氏、寺田清昭氏、ヤマハ株式会社山中晋氏、株式会社モバイルテクノ戸倉綾氏、日本無線株式会社吉田和明氏、高橋義典氏をはじめとする工学院大学数理音響研究室卒業生と関係者の方々に感謝致します。最後に、つねに筆者の心の支えとなり見守って頂き、筆者の夢に賛同し長期間の研究生活を経済的にも支えて頂いた家族に感謝致します。